# Influence Function Based Statistical Inference Under Various Sampling Designs

by

Yi Lu

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
April, 2016

# Abstract

In this dissertation, I consider inference about target parameters under three distinct study designs. The first design, considered in my first paper, is the comprehensive cohort design, whereby patients are first offered enrollment into a randomized trial and if they refuse are offered enrollment into an observational study where they can select their own treatment. In this paper, I develop estimators for two estimands: the comprehensive cohort causal effect and the randomized trial causal effect. The second design, considered in my second paper, is the outcome-dependent two-phase sampling design, where in the first stage inexpensive covariates, treatment choice and outcomes are collected on all patients sampled from a target population and in the second stage expensive covariates, needed to adjust for non-random treatment choice, are collected on a subset of patients. I develop an optimal algorithm for sampling patients at the second stage, where optimality is based on minimizing the asymptotic variance of an estimator of a causal estimand subject to second stage sample size constraints. The third paper applies to any study design where the observed data for individuals can be viewed as independent and identically draws from some distribution. In this paper, we develop a fast double bootstrap procedure for constructing two-sided equal-tailed confidence intervals for estimands which have asymptotically linear estimators. The underlying theme behind all the the papers is the use of influence functions to derive and represent asymptotically linear estimators of target parameters of interest.

## Readers

Daniel O. Scharfstein (Biostatistics)

Elizabeth Ogburn (Biostatistics)

Ellen MacKenzie (Health Policy and Management)

Elizabeth Stuart (Mental Health)

Michael Rosenblum (Alternate, Biostatistics)

Bryan Lau (Alternate, Epidemiology)

# Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Daniel O. Scharfstein, for his guidance, encouragement, and support throughout these years. His enthusiasm and determination in resolving challenging research problems, as well as his high standards and responsibility in mentoring have always been, and will continue to be my source of inspiration in my future career. It would have been impossible to complete this thesis without his help and persistent impetus, given all the difficulties and stresses I have encountered in research and life during my PhD program. I am most fortunate to have worked with him.

I am grateful to my thesis committee members – Drs. Elizabeth Ogburn, Ellen MacKenzie and Elizabeth Stuart, as well as the alternates Michael Rosenblum and Bryan Lau – for their thoughtful comments and constructive suggestions to help me gain deeper insights into different aspects of my dissertation work: science and mathematics, biomedical procedures and statistical methods. In addition, my special thanks go to Dr. Maria Mori Brooks (University of Pittsburgh) for her collaboration on the BARI data analysis for the comprehensive cohort study project; and to Dr. Nicola Lunardon (University of Padua) for his comments and critiques on the fast double bootstrap manuscript. I would also like to thank Drs. Mei-Cheng Wang, Marilyn Albert, Lawrence Wissow, Janet T. Holbrook, Charles Rohde and Xiaobin Wang, who served on my preliminary oral exam committee and have helped broaden my horizons and solidify my research directions during the first few years of my doctoral study.

I feel so honored to have spent the past six years at Johns Hopkins Biostatistics Department, with wonderful faculty and staff. In particular, I thank Drs. Karen Bandeen-Roche, Brian S. Caffo, Marie Diener-West and Hongkai Ji for their encouragement during the pursuit of my doctoral degree; I thank Drs. Thomas A. Louis, Vadim Zipunnikov and Chiung-Yu Huang for sharing their expertise on my research topics; I also thank Mary Joy Argo, Marti Gilbert, Jiong Yang, Mark Chiveral, Debbie Cooper and Debra Moffitt for their help and support on administrative issues.

I want to thank my classmates, colleagues and friends who have been with me all the way and made my experience at Hopkins truly memorable: Francis Abreu, Jiawei Bai, Qing Cai, Detian Deng, Alyssa Frazee, Jonathan Gellar, Fang Han, Bing He, Jeongyong Kim, Shanshan Li, Parichoy Pal Choudhury, Haochang Shou, Yifei Sun, Elizabeth Sweeney, Yenny Webb Vargas, Yingying Wei, Zhenke Wu, Juemin Yang and Yuxin Zhu.

Finally, I would like to dedicate this thesis to my family, especially to my parents Xiaodong Lu and Lifang Xin, whose boundless love and support has always been the tremendous motivation for me to overcome all difficulties and fears, and to pursue my ideals.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In statistical inference based on a random sample of observations, the difference between an estimator and the true value of the target parameter of interest often behaves asymptotically like an average (over observations) of a function of the individual observations. An estimator with this feature is called asymptotically linear and the function is referred to as the influence function (Hampel, 1974). The term influence function is used because to the first order it quantifies the influence of a single observation on the corresponding estimator. Influence functions are useful, as the asymptotic behavior of an asymptotic linear estimator can be approximated, using central limit theory, by considering only its influence function. For example, the asymptotic variance of an estimator can be estimated by an estimate of the variance of the influence function. Furthermore, it is possible to characterize the collection of all influence functions for a target parameter from a statistical model and they can be used to construct asymptotically linear estimators (Tsiatis, 2006).

In this dissertation, we develop statistical methods to address a range of issues related to different sampling designs. The influence function serves as a

common tool used in our statistical developments. The following three sections introduce and motivate the three topics tackled in this thesis.

## 1.1 Causal inference for comprehensive cohort studies

In a comprehensive cohort study of two competing treatments (A,B), clinically eligible individuals are first asked to enroll in a randomized trial and, if they refuse, are then asked to enroll in a parallel observational study in which they can choose treatment according to their own preference. The project is motivated by an ongoing comprehensive cohort study, the FIXIT study, to assess fixation strategies for severe open tibia fractures. The FIXIT study is a multicenter, prospective phase III randomized clinical trial, with patients who refuse randomization being eligible to enroll in a prospective observational study. The study will compare the rate of re-hospitalization for major limb complications under two standard options of treatment: internal fixation with a nail or plate vs. external ring fixation. One of the primary aims is to draw inference using combined information from the randomized trial and the parallel observational study.

We consider estimation of two estimands: (1) comprehensive cohort causal effect – the difference in mean potential outcomes had all patients in the comprehensive cohort received treatment A vs. treatment B and (2) randomized trial causal effect – the difference in mean potential outcomes had all patients enrolled in the randomized trial received treatment A vs. treatment B. We study the class of influence functions under different modeling restrictions in search of efficient and robust estimators. For the comprehensive cohort causal

effect, we introduce two estimators: inverse probability weighted and locally efficient/doubly robust. For the randomized trial causal effect, we introduce four estimators: simple inverse probability weighted, simple robust, enriched doubly robust and locally efficient.

We evaluate finite sample performance of these estimators in a simulation study. We also illustrate our methodology using data from the BARI (Bypass Angioplasty Revascularization Investigation) randomized trial and observational registry to evaluate the effect of percutaneous transluminal coronary balloon angioplasty (PTCA) versus coronary artery bypass grafting (CABG) on 5-year mortality.

## 1.2 Optimal outcome-dependent two-phase sampling

We investigate an optimal outcome-dependent two-phase sampling design to estimate the causal effect of a treatment from observational data. In the first phase, information from all subjects in the study is obtained on treatment, outcome and a set of covariates which are relatively inexpensive to measure. In the second phase, subjects are stratified based on the first phase information, a random subsample of individuals (i.e. validation sample) is drawn from each stratum (with known stratum-specific sampling probabilities), and a rich set of expensive covariates are measured on the validation sample. The causal effect of interest is then estimated based on data collected in both phases.

This topic was motivated by a consulting project in which a pharmaceutical company was interested in comparing the efficacy of their drug to a competitor's drug based on data from a very large healthcare database. The problem was

that the key covariates (e.g., co-morbidities) needed to adjust for confounding required an expensive medical record review process, whereas treatment, outcome and basic covariates (e.g., age, gender) were readily available. The pharmaceutical company wanted to know how to select patients for medical record review.

Given the first phase data and a specified estimator of the causal effect, we propose an optimal choice for the stratum-specific sampling probabilities in the second phase by minimizing the variance of the estimator given validation sample size constraints. We consider four estimators of the causal effect: simple inverse probability weighted and enriched doubly robust each with sampling probabilities known and with sampling probabilities estimated. We investigate the general form of the influence functions for these estimators, to capture the relationship between the within strata sampling fractions and the asymptotic variance of the resulting estimator.

We present a detailed simulation study, designed to evaluate the finite sample performance of these estimators under different validation sampling schemes. We also illustrate our methods using data from a large observational study to estimate the causal effect of right heart catheterization (RHC) on 30-day survival under various hypothetical two-phase sampling schemes.

## 1.3 Influence function based fast double bootstrap confidence intervals

For small to medium sized samples, equal-tailed two-sided confidence intervals based on asymptotic normal approximation or bootstrap methods such as the percentile bootstrap and studentized bootstrap, often fail to provide accurate

coverage. Iterated bootstrap methods have been proposed to reduce the coverage error by calibrating the nominal level using nested levels of resampling, which in practice can be extremely computationally intensive.

This topic was motivated by a project where sensitivity analysis methods were being developed for estimating treatment effects from clinical trials in the presence of potentially informative missing data. In simulations, Wald-based confidence intervals based on influence functions and single layer bootstrap confidence intervals were seen to have poor coverage in small to moderate sized samples. A double bootstrap solution was considered to be computationally intensive. A natural question is whether there is a fast alternative. This paper is a first step at addressing this problem.

Specifically, we propose a fast double bootstrap procedure for constructing confidence intervals for a parameter whose estimator is defined as the solution to an estimating function involving nuisance parameters. Our resampling process directly exploits the influence function representation, and a saddle-point approximation approach is employed to avoid inner-level resamplings. The resulting equal-tailed two-sided confidence intervals are shown to be at least third order accurate, i.e. better coverage accuracy than single layer bootstrap intervals, yet without extra computational burden. A simulation study demonstrates that, compared to other methods, our fast double bootstrap interval has desirable empirical coverage for small to medium sample sizes.

## 1.4 Overview of dissertation

This dissertation is organized as follows. Chapters 2-4 provide the details of the topics described in Sections 1.1-1.3 above. In each chapter, we present (i) an introduction to the statistical problem, review of the existing statistical methodology, and the relevance of our contribution; (ii) a description of the theoretical framework and our statistical methods; (iii) numerical results from simulation study and/or real data analysis; and (iv) a summary of conclusions, with a discussion of limitations and future research directions. Chapter 5 is devoted to concluding remarks.

# Chapter 2

# Causal Inference for Comprehensive Cohort Studies

## 2.1 Introduction

Randomized controlled trials (RCTs) are considered to be the gold standard in evaluating the effect of treatments, primarily because the randomization process probabilistically ensures that treatment groups are balanced with respect to measured and unmeasured prognostic factors. A well conducted RCT is said to have high internal validity. However, its external validity (i.e., generalizability of results to a broader population) is not guaranteed. This is because eligible patients who agree to enroll in an RCT may be not be a representative sample of all eligible patients. Due to the the tension between internal and external validity, researchers have recommended that all clinically eligible patients, agreeing to randomization or not, should be enrolled and studied (Fielding et al., 1999).

Olschewski and Scheurlen (1985) introduced the comprehensive cohort study (CCS) design for evaluating competing treatments (say, A and B) in which clinically eligible participants are first asked to enroll in a randomized trial and, if they refuse, are then asked to enroll in a parallel observational study (OBS) in

which they can choose treatment according to their own preference (see Figure 2.1(a)). Since the CCS design incorporates patient preference, it has also been referred to as a (partially randomized) patient preference trial (Brewin and Bradley, 1989; Brocklehurst, 1997; Torgerson and Sibbald, 1998).

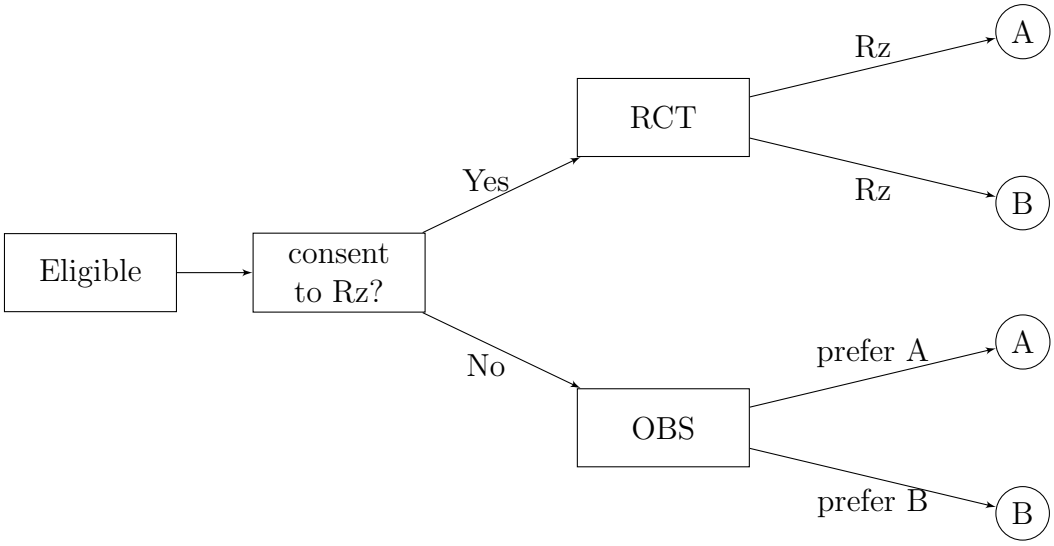
King et al. (2005) conducted a systematic review of randomized trials incorporating participants' preferences published during 1966-2004. In addition to the CCS design, they examined the two-stage randomized (TSR) design proposed by Rücker (1989) and Wennberg et al. (1993) (see Figure 2.1(b)). In contrast to the CCS design, the TSR design has a first stage randomization into either the RCT or OBS. Of the 32 trials identified by King et al. (2005), 27 (84%) were CCS designs.

In this paper, we focus on inference about treatment effects from a CCS, where the primary outcome (continuous or binary) is to be measured at a fixed point in time after treatment assignment. In particular, we are interested in drawing inference about two causal estimands: *comprehensive cohort causal effect* – the difference in mean potential outcomes had all patients in the comprehensive cohort received treatment A vs. treatment B, and *randomized trial causal effect* – the difference in mean potential outcomes had all patients enrolled in the randomized trial received treatment A vs. treatment B. The first estimand has greater external validity than the second. Our goal is to use all the available data to draw inference about these estimands.

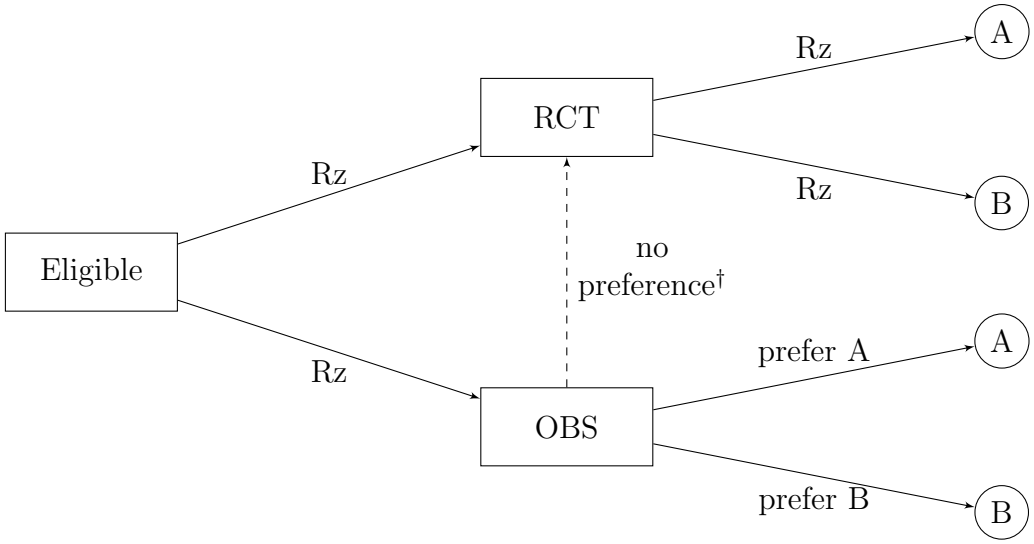
Marcus (1997) has investigated the difference between the two estimands in a CCS design, called "non-consent bias", and suggested adjusting for baseline covariates to account for the difference between the RCT population and the comprehensive cohort. Otherwise, most of the literature that describes methods

for analyzing treatments effects from a CCS doesn't directly address our causal objectives. Rather, the main focus is on estimating treatment effects separately for the RCT and OBS (see, for example, Olschewski et al. (1992), Henshaw et al. (1993), Nicolaides et al. (1994), Schmoor et al. (1996), King et al. (1997), Detre et al. (1999), Bedi et al. (2000), Brooks et al. (2000), Kerry et al. (2000), King et al. (2000), Rovers et al. (2001), Jensen et al. (2003), Schmoor et al. (2008)), with some adjusting for confounding in the OBS. With the exception of Olschewski et al. (1992), King et al. (1997) and Brooks et al. (2000), there is no borrowing of information between the RCT and OBS.

In this paper, we derive, using semiparametric theory (Tsiatis, 2006), estimators of the comprehensive cohort and randomized trial causal effects under a set of unconfoundedness assumptions regarding randomization consent status and treatment assignment. In Section 2.2, we introduce notation, data structure, and main assumptions. In Section 2.3, we propose two estimators for the comprehensive cohort causal effect: inverse probability weighted (CC-IPW), locally efficient and doubly robust (CC-LEDR); and four estimators for the randomized trial causal effect: simple inverse probability weighted (RCT-SIPW), simple robust (RCT-SR), enriched doubly robust (RCT-EDR) and locally efficient (RCT-LE). Section 2.4 presents a simulation study that evaluates the finite sample performance of our proposed estimators. In Section 2.5, we illustrate our methods using data from the BARI (Bypass Angioplasty Revascularization Investigation) randomized trial and observational registry to evaluate the effect of percutaneous transluminal coronary balloon angioplasty (PTCA) versus coronary artery bypass grafting (CABG) on 5-year mortality. The last section is devoted to a discussion.




(a) Comprehensive cohort study (CCS) design




(b) Two-stage randomized (TSR) design (†In Rücker (1989) design, participants randomized to OBS in the first stage who do not have a strong preference for a treatment are randomized to a treatment.)

Figure 2.1: Flowchart of CSS and TSR study designs for comparing two treatments A vs. B (Rz: randomization)

## 2.2 Notation and framework

Let $X$ denote a vector of baseline covariates and let $Y$ be the observed outcome (continuous or binary). Let $R$ denote the randomization consent indicator (1 for RCT, 0 for OBS) and let $T$ denote the treatment assignment indicator (1 for treatment A, 0 for treatment B). The observed data for an individual are $O = (X', Y, R, T)'$. We assume that we observe $n$ independent and identically distributed copies of $O$. Let $Y_1$ and $Y_0$ denote an eligible patient's potential outcome under treatment A and B, respectively.

Our goal is to use the observed data to draw inference about the comprehensive cohort causal effect:

$$\Delta_{CC} = \underbrace{E[Y_1]}_{\mu_1^*} - \underbrace{E[Y_0]}_{\mu_0^*}$$

and the randomized trial causal effect:

$$\Delta_{RCT} = \underbrace{E[Y_1|R=1]}_{\mu_{11}^*} - \underbrace{E[Y_0|R=1]}_{\mu_{10}^*}$$

To identify $\Delta_{CC}$ and $\Delta_{RCT}$ from the observed data, we posit assumptions sufficient for identification of $\mu_t^*$ and $\mu_{1t}^*$ $(t = 0, 1)$. We make the *stable unit treatment value assumption* as discussed in Rubin (1986), which states that the potential outcomes of a patient is unaffected by the randomization consent and treatment decision of any other patient. We make the *consistency* assumption that states that the observed outcome under the treatment actually received is equal to the potential outcome under that treatment (i.e., $Y = TY_1 + (1-T)Y_0$). Further, we assume

(A1). *Conditional unconfoundedness of consent to randomization*:

$$R \perp (Y_1, Y_0) | \, X$$

(A2). *Unconfoundedness of treatment assignment within the RCT*:

$$T \perp (Y_1, Y_0, X) | \, R = 1$$

(A3). *Conditional unconfoundedness of treatment selection within the OBS*:

$$T \perp (Y_1, Y_0) | \, R = 0, X$$

Assumption (A1) implies that, conditional on $(X, Y_1, Y_0)$, consent to randomization is a Bernoulli process with the consent probability $\lambda^*(X) = P\left[R = 1 | \, X\right]$, i.e. only depending upon the covariates $X$. Assumptions (A2) and (A3) indicate that, conditional on $(X, Y_1, Y_0, R)$, treatment selection is a Bernoulli process with the selection probability $\pi_t^*(R, X) = P\left[T = t | \, R, X\right]$ and does not depend on $X$ when $R = 1$. We further assume that $\lambda^*(X)$ and $\pi_t^*(R, X)$ are strictly greater than 0 and less than 1 for all $X$ and $t$.

For convenience, see Table 2.1 for a complete list of our notation, with some additional symbols used in the estimation section.

## 2.3 Estimation

Our overarching estimation strategy proceeds by first finding unbiased estimating functions for $\mu_t^*$ and $\mu_{1t}^*$. These estimating functions depend on nuisance functions. We propose estimators for $\mu_t^*$ and $\mu_{1t}^*$ by solving empirical versions of these estimating functions with estimators plugged-in for the nuisance functions. We also determine the asymptotic properties of our proposed estimators.

Table 2.1: Notation

| Symbol | Description ($t = 0, 1.$) |
|---|---|
| $R$ | Randomization consent indicator (1 for RCT, 0 for OBS) |
| $T$ | Treatment indicator (1 for A, 0 for B) |
| $X$ | Baseline covariates |
| $Y_t$ | Potential outcome under treatment $t$ |
| $Y$ | Observed outcome |
| $\mu_t^*$ | $E[Y_t]$ |
| $\mu_{1t}^*$ | $E[Y_t \vert R = 1]$ |
| $\mu_t^*(X)$ | $E[Y_t \vert X]$ |
| $\Delta_{CC}$ | $\mu_1^* - \mu_0^*$ |
| $\Delta_{RCT}$ | $\mu_{11}^* - \mu_{10}^*$ |
| $\lambda^*$ | $P[R = 1]$ |
| $\lambda^*(X)$ | $P[R = 1 \vert X]$ |
| $\pi_t^*(R, X)$ | $P\left[T = t \vert\, R, X\right]$ |
| $\rho_t^*$ | $P\left[T = t, R = 1\right]$ |

### 2.3.1 Inverse probability weighted estimators

Consider the following the inverse probability weighted estimating functions:

$$U_t^{\text{CC-IPW}}\{O; \mu_t, \pi_t(R, X)\} = \frac{I\{T = t\}}{\pi_t(R, X)}(Y - \mu_t) \tag{2.1}$$

$$U_{1t}^{\text{RCT-SIPW}}(O; \mu_{1t}, \rho_t) = \frac{I\{T = t\}R}{\rho_t}(Y - \mu_{1t}) \tag{2.2}$$

for $\mu_t^*$ and $\mu_{1t}^*$, respectively. We label (2.1) as the *inverse probability weighted* (IPW) estimating function, and (2.2) as the *simple inverse probability weighted* (SIPW) estimating function, where the word "simple" is added to reflect the use of information only from the RCT. Under our assumptions, both functions can be shown to have mean 0 (i.e., unbiased) when evaluated at the truth $(\mu_t^*, \pi_t^*(R, X))$ and $(\mu_{1t}^*, \rho_t^*)$, respectively.

**CC-IPW**

In order to draw $\sqrt{n}-$inference about $\mu_t^*$ based on (2.1) , we need an estimator of $\pi_t^*(R,X)$ converging at a rate faster than $n^{1/4}$. To proceed, we consider a fully parametric model for $\pi_t^*(R,X) = \pi_t(R,X;\alpha^*)$, where $\alpha^*$ denotes the true value of the model parameter (vector) $\alpha$. We shall assume the following logistic model

$$\begin{aligned} \text{logit}\left\{\pi_1(R,X;\alpha^*)\right\} &= l(R,X;\alpha^*) \\ &= R\alpha_1^* + (1-R)l_0(X;\alpha_0^*) \qquad (\mathcal{M}_\pi) \end{aligned}$$

where $\alpha_1$ is a scalar parameter, $l_0(X;\alpha_0)$ is a specified function of $X$ and the parameter vector $\alpha_0$, $\alpha_1^*$ and $\alpha_0^*$ are the true values of $\alpha_1$ and $\alpha_0$, respectively, and $\alpha^* = (\alpha_1^*, \alpha_0^{*\prime})'$. Under our assumptions, Model $\mathcal{M}_\pi$ implies

$$\pi_t(R,X;\alpha^*) = R\pi_t(1;\alpha_1^*) + (1-R)\pi_t(0,X;\alpha_0^*)$$

where

$$\pi_t(1;\alpha_1^*) = \frac{\exp(t\alpha_1^*)}{1+\exp(\alpha_1^*)}$$

$$\pi_t(0,X;\alpha_0^*) = \frac{\exp\left\{tl_0(X;\alpha_0^*)\right\}}{1+\exp\left\{l_0(X;\alpha_0^*)\right\}}$$

Based on this logistic model $\mathcal{M}_\pi$ , $\alpha^*$ can be estimated as the solution $\hat{\alpha}$ to

$$E_n\left[S_\alpha(T,R,X;\alpha)\right] = 0,$$

where $E_n[\cdot]$ is the empirical expectation operator and

$$S_\alpha(T,R,X;\alpha) = \frac{\partial l(R,X;\alpha)}{\partial \alpha}\left\{T - \pi_1(R,X;\alpha)\right\}$$

is the logistic regression score function. Then we can estimate $\mu_t^*$ by solving

$$E_n\left[U_t{}^{\text{CC-IPW}}(O;\mu_t,\hat{\alpha})\right]=0, \tag{2.3}$$

where

$$U_t^{\text{CC-IPW}}(O;\mu_t,\alpha)=\frac{I\{T=t\}}{\pi_t(R,X;\alpha)}(Y-\mu_t)$$

has mean 0 when evaluated at $(\mu_t^*,\alpha^*)$. Let $\hat{\mu}_t^{\text{CC-IPW}}$ denote the solution to (2.3). Under correct specification of model $\mathcal{M}_\pi$ and some additional regularity conditions, it can be shown that $\hat{\mu}_t^{\text{IPW}}$ is a *regular and asymptotically linear* (RAL) estimator of $\mu_t^*$, with influence function

$$\begin{aligned}\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)&=U_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\\&\quad-E\left[\frac{\partial U_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)}{\partial\alpha'}\right]E\left[\frac{\partial S_\alpha(T,R,X;\alpha^*)}{\partial\alpha'}\right]^{-1}\\&\quad\times S_\alpha(T,R,X;\alpha^*)\end{aligned}$$

Hence, the asymptotic variance of $\hat{\mu}_t^{\text{CC-IPW}}$ is equal to $E[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)^2]$, which can be estimated by $E_n[\widehat{\text{IF}}_t^{\text{CC-IPW}}(O;\hat{\mu}_t^{\text{CC-IPW}},\hat{\alpha})^2]$, with $\widehat{\text{IF}}_t^{\text{CC-IPW}}(O;\mu_t,\alpha)$ the same as $\text{IF}_t^{\text{CC-IPW}}(O;\mu_t,\alpha)$ except that the expectations are replaced with empirical averages.

**RCT-SIPW**

To draw inference about $\mu_{1t}^*$ based on (2.2), we estimate $\rho_t^*=P\left[T=t,R=1\right]$ by

$$\hat{\rho}_t:=E_n\left[I\{T=t,R=1\}\right]$$

We then estimate $\mu_{1t}^*$ as the solution $\hat{\mu}_{1t}^{\text{RCT-SIPW}}$ to

$$E_n\left[U_{1t}^{\text{RCT-SIPW}}(O;\mu_{1t},\hat{\rho}_t)\right]=0,$$

The influence function for $\hat{\mu}_{1t}^{\text{RCT-SIPW}}$ is $U_{1t}^{\text{RCT-SIPW}}(O; \mu_{1t}^*, \rho_t^*)$, and we can estimate its asymptotic variance by $E_n[U_{1t}^{\text{RCT-SIPW}}(O; \hat{\mu}_{1t}^{\text{RCT-SIPW}}, \hat{\rho}_t)^2]$.

### 2.3.2 Alternative estimators under $\mathcal{M}_\pi$

The inverse probability weighted estimators discussed above are not efficient. Based upon the semiparametric theory of inference for coarsened data (Tsiatis, 2006), we can derive the most efficient influence function (corresponding to the RAL estimator with the smallest variance) by projecting the influence function of any RAL estimator onto the semiparametric tangent space.

We consider the semiparametric model $\mathcal{P}_\pi$ which imposes restrictions on the distribution of the observed data through assumptions (A1) – (A3) and model $\mathcal{M}_\pi$. The semiparametric tangent space for this model ($\mathscr{T}_\pi^O$) is defined as the mean square closure of the linear space of score functions from all parametric submodels contained in the model. To proceed with the desired projection, we show in Appendix I that the orthogonal complement of $\mathscr{T}_\pi^O$ has the form:

$$\mathscr{T}_\pi^{O,\perp} = \Lambda_h \oplus \Lambda_b \oplus \Lambda_a, \tag{2.4}$$

where $\oplus$ denotes the direct sum and

$$\Lambda_h = \Lambda_{h_0} \oplus \Lambda_{h_1}$$

$$\Lambda_{h_\tau} = \{\phi_{2\tau}\{O; \alpha^*, \lambda^*(X), h_\tau\} : E\left[h_\tau(X, Y_\tau) \middle| X\right] = 0\} \quad \tau = 0, 1.$$

$$\Lambda_b = \left\{\phi_3\{O; \alpha_0^*, \lambda^*(X), b\} : E\left[\frac{\partial l_0(X; \alpha_0^*)}{\partial \alpha_0} b(X)\right] = 0\right\}$$

$$\Lambda_a = \{\phi_4\{O; \alpha_1^*, \lambda^*(X), a\} : E\left[a(X)\right] = 0\}$$

with

$$\phi_{2\tau}\{O;\alpha^*,\lambda^*(X),h_\tau\} = \left\{\frac{I\{T=\tau\}R}{\pi_\tau(1;\alpha_1^*)\lambda^*(X)} - \frac{I\{T=\tau\}(1-R)}{\pi_\tau(0,X;\alpha_0^*)\{1-\lambda^*(X)\}}\right\} h_\tau(X,Y_\tau)$$

$$\phi_3\{O;\alpha_0^*,\lambda^*(X),b\} = \frac{1-R}{1-\lambda^*(X)}\frac{T-\pi_1(0,X;\alpha_0^*)}{\pi_1(0,X;\alpha_0^*)\pi_0(0,X;\alpha_0^*)}b(X)$$

$$\phi_4\{O;\alpha_1^*,\lambda^*(X),a\} = \frac{R}{\lambda^*(X)}\frac{T-\pi_1(1;\alpha_1^*)}{\pi_1(1;\alpha_1^*)\pi_0(1;\alpha_1^*)}a(X).$$

Importantly, $\Lambda_{h_0}, \Lambda_{h_1}, \Lambda_b, \Lambda_a$ are pairwise orthogonal to each other. Appendix III shows how to project an arbitrary function of the observed data (with mean zero and finite variance) onto each of these spaces. Based on these results, we can derive the projection of the influence functions for CC-IPW and CC-SIPW estimators onto the semiparametric tangent space $\mathscr{T}_\pi^O$, for more efficient estimators of $\mu_t^*$ and $\mu_{1t}^*$, respectively.

**Efficient and robust estimator for $\mu_t^*$ (CC-LEDR)**

As mentioned above, the most efficient influence function for estimating $\mu_t^*$ under the semiparametric model $\mathcal{P}_\pi$ is the projection

$$
\begin{aligned}
&\phi_t^{\text{CC-LEDR}}\{O;\mu_t^*,\alpha^*,\lambda^*(X),\mu_t^*(X)\}\\
&:=\Pi\big[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\mathscr{T}_\pi^O\big]\\
&=\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)-\Pi\big[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\mathscr{T}_\pi^{O,\perp}\big]\\
&=\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)-\Pi\big[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\Lambda_{h_0}\big]\\
&\quad-\Pi\big[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\Lambda_{h_1}\big]-\Pi\big[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\Lambda_b\big]\\
&\quad-\Pi\big[\text{IF}_t^{\text{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\Lambda_a\big]\\
&=\frac{I\{T=t\}}{\pi_t(R,X;\alpha^*)}(Y-\mu_t^*)+(-1)^t\left\{\frac{T-\pi_1(R,X;\alpha^*)}{\pi_t(R,X;\alpha^*)}\right\}\{\mu_t^*(X)-\mu_t^*\}\\
&\quad-\left\{\frac{I\{T=t\}R}{\pi_t(1;\alpha_1^*)\lambda^*(X)}-\frac{I\{T=t\}(1-R)}{\pi_t(0,X;\alpha_0^*)\{1-\lambda^*(X)\}}\right\}\{Y-\mu_t^*(X)\}\\
&\qquad\times\left\{\frac{1}{\pi_t(1;\alpha_1^*)\lambda^*(X)}+\frac{1}{\pi_t(0,X;\alpha_0^*)\{1-\lambda^*(X)\}}\right\}^{-1}\left\{\frac{1}{\pi_t(1;\alpha_1^*)}-\frac{1}{\pi_t(0,X;\alpha_0^*)}\right\},
\end{aligned}
\tag{2.5}
$$

where $\Pi[\cdot|\cdot]$ is the projection operator. Treated as an estimating function, (2.5) has mean 0 when evaluated at $\{\mu_t^*,\alpha^*,\lambda^*(X),\mu_t^*(X)\}$.

To draw $\sqrt{n}-$inference about $\mu_t^*$, we consider parametric models: $\lambda^*(X)=\lambda(X;\gamma^*)$ and $\mu_t^*(X)=\mu_t(X;\eta_t^*)$, where $\gamma^*,\eta_t^*$ denote the true values of the

model parameter vectors $\gamma, \eta_t$ respectively. For the randomization consent indicator, we posit the logistic model

$$\text{logit}\left\{\lambda(X;\gamma^*)\right\} = q(X;\gamma^*) \qquad (\mathcal{M}_\lambda)$$

where $q(X;\gamma^*)$ is a specified function of $X$ and $\gamma$. In this model, $\gamma^*$ can be estimated as the solution $\hat{\gamma}$ to

$$E_n\left[S_\gamma(R, X;\gamma)\right] = 0,$$

where

$$S_\gamma(R, X;\gamma) = \frac{\partial q(X;\gamma)}{\partial \gamma}\left\{R - \lambda(X;\gamma)\right\}$$

is the associated score function. The parameter $\eta_t^*$ can be estimated using an estimating function of the form:

$$I\{T = t\}\underbrace{a_t(X;\eta_t)\{Y - \mu_t(X;\eta_t)\}}_{S_{\eta_t}(Y,X;\eta_t)}$$

where $a_t(X;\eta_t)$ is a specified function of $X$ and $\eta_t$ that is of the same dimension as $\eta_t$. Note that this estimating function has mean zero at $\eta_t^*$ because assumptions (A1) – (A3) imply that

$$T \perp (Y_1, Y_0) |\, X$$

We estimate $\eta_t^*$ as the solution $\hat{\eta}_t$ to

$$E_n\left[I\{T = t\}S_{\eta_t}(Y, X;\eta_t)\right] = 0.$$

Consequently, we may replace $\alpha^*$, $\lambda^*(X), \mu_t^*(X)$ in (2.5) with $\hat{\alpha}$, $\lambda(X;\hat{\gamma})$ and $\mu_t(X;\hat{\eta}_t)$ respectively, and define the estimator $\hat{\mu}_t^{\text{CC}-\text{LEDR}}$ as the solution to

$$E_n\left[\phi_t^{\text{CC}-\text{LEDR}}\{O;\mu_t,\hat{\alpha},\lambda(X;\hat{\gamma}),\mu_t(X;\hat{\eta}_t)\}\right] = 0.$$

We name (2.5) the *locally efficient doubly robust* (LEDR) influence function, for the following reasons. First, it's not only the most efficient influence function under the restrictions of assumptions (A1) – (A3) and model $\mathcal{M}_\pi$, but also the most efficient under the additional model restriction $\mathcal{M}_\lambda$, which will be discussed later in Section 2.3.3. Second, the influence function (2.5) has the following properties:

(i). $E\left[\phi_t^{\text{CC}-\text{LEDR}}\{O; \mu_t^*, \alpha^*, \lambda(X;\gamma^*), \mu_t(X;\tilde{\eta}_t)\}\right] = 0$ for any $\tilde{\eta}_t$;

(ii). $E\left[\phi_t^{\text{CC}-\text{LEDR}}\{O; \mu_t^*, \tilde{\alpha}, \lambda(X;\tilde{\gamma}), \mu_t(X;\eta_t^*)\}\right] = 0$ for any $\tilde{\alpha}, \tilde{\gamma}$;

and thus is said to be "doubly robust" in the sense that the resulting estimator $\hat{\mu}_t^{\text{CC}-\text{LEDR}}$ will be consistent and asymptotically normal when either the joint model $(\mathcal{M}_\pi \cap \mathcal{M}_\lambda)$ or the model for $\mu_t^*(X)$ is correctly specified. As a result, the influence function for $\hat{\mu}_t^{\text{CC}-\text{LEDR}}$ can be shown to be

$$
\begin{aligned}
&\text{IF}_t^{\text{CC-LEDR}}(O; \mu_t^*, \tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}_t) \\
&= \phi_t^{\text{CC-LEDR}}\{O; \mu_t^*, \tilde{\alpha}, \lambda(X;\tilde{\gamma}), \mu_t(X;\tilde{\eta}_t)\} \\
&\quad - E\left[\frac{\partial \phi_t^{\text{CC-LEDR}}\{O; \mu_t^*, \tilde{\alpha}, \lambda(X;\tilde{\gamma}), \mu_t(X;\tilde{\eta}_t)\}}{\partial \alpha'}\right] E\left[\frac{\partial S_\alpha(T, R, X; \tilde{\alpha})}{\partial \alpha'}\right]^{-1} \\
&\qquad \times S_\alpha(T, R, X; \tilde{\alpha}) \\
&\quad - E\left[\frac{\partial \phi_t^{\text{CC-LEDR}}\{O; \mu_t^*, \tilde{\alpha}, \lambda(X;\tilde{\gamma}), \mu_t(X;\tilde{\eta}_t)\}}{\partial \gamma'}\right] E\left[\frac{\partial S_\gamma(R, X; \tilde{\gamma})}{\partial \gamma'}\right]^{-1} S_\gamma(R, X; \tilde{\gamma}) \\
&\quad - E\left[\frac{\partial \phi_t^{\text{CC-LEDR}}\{O; \mu_t^*, \tilde{\alpha}, \lambda(X;\tilde{\gamma}), \mu_t(X;\tilde{\eta}_t)\}}{\partial \eta_t'}\right] E\left[I\{T=t\}\frac{\partial S_{\eta_t}(Y, X; \tilde{\eta}_t)}{\partial \eta_t'}\right]^{-1} \\
&\qquad \times I\{T=t\} S_{\eta_t}(Y, X; \tilde{\eta}_t),
\end{aligned}
\tag{2.6}
$$

where either $(\tilde{\alpha}', \tilde{\gamma}')' = (\alpha^{*\prime}, \gamma^{*\prime})'$ (the models for $\pi_t^*(R, X)$ and $\lambda^*(X)$ are correctly specified) or $\tilde{\eta}_t = \eta_t^*$ (the model for $\mu_t^*(X)$ is correctly specified). Further, the second and third terms on the right hand side of equation (2.6) vanish if $\tilde{\eta}_t = \eta_t^*$, while the fourth term vanishes if $(\tilde{\alpha}', \tilde{\gamma}')' = (\alpha^{*\prime}, \gamma^{*\prime})'$. The asymptotic variance of $\hat{\mu}_t^{\text{CC}-\text{LEDR}}$ based on (2.6) is $E[\text{IF}_t^{\text{CC-LEDR}}(O; \mu_t^*, \tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}_t)^2]$, which is referred to as the "doubly robust variance" for similar reasons; we estimate it by $E_n[\widehat{\text{IF}}_t^{\text{CC}-\text{LEDR}}(O; \hat{\mu}_t^{\text{CC}-\text{LEDR}}, \hat{\alpha}, \hat{\gamma}, \hat{\eta}_t)^2]$, where $\widehat{\text{IF}}_t^{\text{CC}-\text{LEDR}}(O; \mu_t, \alpha, \gamma, \eta_t)$ is the same as $\text{IF}_t^{\text{CC}-\text{LEDR}}(O; \mu_t, \alpha, \gamma, \eta_t)$ except that the expectations are replaced by empirical averages.

**Enriched doubly-robust estimator of $\mu_{1t}^*$ (RCT-EDR)**

Using semiparametric theory, the most efficient influence function for estimating $\mu_{1t}^*$ under the semiparametric model $\mathcal{P}_\pi$ is the projection

$$\begin{aligned}
&\phi_{1t}^{\text{RCT}-\text{EDR}}\{O;\mu_{1t}^*,\rho_t^*,\alpha^*,\lambda^*(X),\mu_t^*(X)\} \\
&:= \Pi\big[U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\mathscr{T}_\pi^O\big] \\
&= U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*) - \Pi\big[U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\mathscr{T}_\pi^{O,\perp}\big] \\
&= U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*) - \Pi\big[U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_{h_0}\big] \\
&\quad - \Pi\big[U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_{h_1}\big] - \Pi\big[U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_b\big] \\
&\quad - \Pi\big[U_{1t}^{\text{RCT}-\text{SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_a\big] \\
&= \frac{I\{T=t\}R}{\rho_t^*}(Y-\mu_{1t}^*) + (-1)^t R\left\{\frac{T-\pi_1(1;\alpha_1^*)}{\rho_t^*}\right\}\{\mu_t^*(X)-\mu_{1t}^*\} \\
&\quad - \left\{\frac{I\{T=t\}R}{\pi_t(1;\alpha_1^*)\lambda^*(X)} - \frac{I\{T=t\}(1-R)}{\pi_t(0,X;\alpha_0^*)\{1-\lambda^*(X)\}}\right\}\left\{\frac{Y-\mu_t^*(X)}{\rho_t^*}\right\} \\
&\quad \times \left\{\frac{1}{\pi_t(1;\alpha_1^*)\lambda^*(X)} + \frac{1}{\pi_t(0,X;\alpha_0^*)\{1-\lambda^*(X)\}}\right\}^{-1}, \qquad (2.7)
\end{aligned}$$

Treated as an estimating function, (2.7) has mean zero when evaluated at the truth $\{\mu_{1t}^*,\rho_t^*,\alpha^*,\lambda^*(X),\mu_t^*(X)\}$. As above, we employ parametric models $\lambda^*(X)=\lambda(X;\gamma^*)$ and $\mu_t^*(X)=\mu_t(X;\eta_t^*)$. We obtain the RAL estimator $\hat{\mu}_{1t}^{\text{RCT}-\text{EDR}}$ as the solution to

$$E_n\left[\phi_{1t}^{\text{RCT}-\text{EDR}}\{O;\mu_{1t},\hat{\rho}_t,\hat{\alpha},\lambda(X;\hat{\gamma}),\mu_t(X;\hat{\eta}_t)\}\right]=0,$$

with the estimators $\hat{\rho}_t,\hat{\alpha},\hat{\gamma},\hat{\eta}_t$ defined previously.

We call (2.7) the *enriched doubly robust* (EDR) influence function, with the word "enriched" to emphasize the usage of additional information from the OBS. This is in contrast to $\hat{\mu}_{1t}^{\text{RCT}-\text{SIPW}}$ which only used data from the RCT. In addition, the influence function (2.7) also has the same "doubly robustness" property of CC-LEDR influence function (2.5). That is, $\hat{\mu}_{1t}^{\text{RCT}-\text{EDR}}$ will be consistent and asymptotically normal when either the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$) or the model for $\mu_t^*(X)$ is correctly specified; accordingly, the influence function for $\hat{\mu}_{1t}^{\text{RCT}-\text{EDR}}$ is

$$
\begin{aligned}
&\text{IF}_{1t}^{\text{RCT-EDR}}(O; \mu_{1t}^*, \rho_t^*, \tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}_t) \\
&= \phi_{1t}^{\text{RCT-EDR}}\{O; \mu_{1t}^*, \rho_t^*, \tilde{\alpha}, \lambda(X; \tilde{\gamma}), \mu_t(X; \tilde{\eta}_t)\} \\
&\quad - E\left[\frac{\partial \phi_{1t}^{\text{RCT-EDR}}\{O; \mu_{1t}^*, \rho_t^*, \tilde{\alpha}, \lambda(X; \tilde{\gamma}), \mu_t(X; \tilde{\eta}_t)\}}{\partial \alpha'}\right] E\left[\frac{\partial S_\alpha(T, R, X; \tilde{\alpha})}{\partial \alpha'}\right]^{-1} \\
&\qquad \times S_\alpha(T, R, X; \tilde{\alpha}) \\
&\quad - E\left[\frac{\partial \phi_{1t}^{\text{RCT-EDR}}\{O; \mu_{1t}^*, \rho_t^*, \tilde{\alpha}, \lambda(X; \tilde{\gamma}), \mu_t(X; \tilde{\eta}_t)\}}{\partial \gamma'}\right] E\left[\frac{\partial S_\gamma(R, X; \tilde{\gamma})}{\partial \gamma'}\right]^{-1} S_\gamma(R, X; \tilde{\gamma}) \\
&\quad - E\left[\frac{\partial \phi_{1t}^{\text{RCT-EDR}}\{O; \mu_{1t}^*, \rho_t^*, \tilde{\alpha}, \lambda(X; \tilde{\gamma}), \mu_t(X; \tilde{\eta}_t)\}}{\partial \eta_t'}\right] E\left[I\{T = t\}\frac{\partial S_{\eta_t}(Y, X; \tilde{\eta}_t)}{\partial \eta_t'}\right]^{-1} \\
&\qquad \times I\{T = t\} S_{\eta_t}(Y, X; \tilde{\eta}_t),
\end{aligned}
\tag{2.8}
$$

where either $(\tilde{\alpha}', \tilde{\gamma}')' = (\alpha^{*\prime}, \gamma^{*\prime})'$ (the models for $\pi_t^*(R, X)$ and $\lambda^*(X)$ are correctly specified) or $\tilde{\eta}_t = \eta_t^*$ (the model for $\mu_t^*(X)$ is correctly specified). Again, on the right hand side of equation (2.8), the second and third terms vanish if $\tilde{\eta}_t = \eta_t^*$, while the fourth term vanishes if $(\tilde{\alpha}', \tilde{\gamma}')' = (\alpha^{*\prime}, \gamma^{*\prime})'$. Similarly, the "doubly robust" asymptotic variance of $\hat{\mu}_{1t}^{\text{RCT}-\text{EDR}}$ is $E[\text{IF}_{1t}^{\text{RCT}-\text{EDR}}(O; \mu_{1t}^*, \rho_t^*, \tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}_t)^2]$,

which can be estimated by $E_n[\widehat{\text{IF}}_{1t}^{\text{RCT-EDR}}(O; \hat{\mu}_{1t}^{\text{RCT-EDR}}, \hat{\rho}_t, \hat{\alpha}, \hat{\gamma}, \hat{\eta}_t)^2]$, where $\widehat{\text{IF}}_{1t}^{\text{RCT-EDR}}(O; \mu_{1t}, \rho_t, \alpha, \gamma, \eta_t)$ is the same as $\text{IF}_{1t}^{\text{RCT-EDR}}(O; \mu_{1t}, \rho_t, \alpha, \gamma, \eta_t)$ except that the expectations are replaced by empirical averages.

**Simple robust estimator of $\mu_{1t}^*$ (RCT-SR)**

When we examine the RCT-EDR estimating function (2.7), we notice that the first part is itself an influence function

$$\begin{aligned}&\phi_{1t}^{\text{RCT-SR}}\{O; \mu_{1t}^*, \rho_t^*, \alpha_1^*, \mu_t^*(X)\}\\ &:= \frac{I\{T=t\}R}{\rho_t^*}(Y - \mu_{1t}^*) + (-1)^t R\left\{\frac{T - \pi_1(1;\alpha_1^*)}{\rho_t^*}\right\}\{\mu_t^*(X) - \mu_{1t}^*\}, \qquad (2.9)\end{aligned}$$

This influence function takes the form of an augmented inverse probability weighted estimating function and is, according to the theory in Tsiatis (2006), the most efficient influence function under assumption (A2) with only RCT data. In addition, the influence function (2.9) has the following robustness property:

$$E\left[\phi_{1t}^{\text{RCT-SR}}\{O; \mu_{1t}^*, \rho_t^*, \alpha_1^*, f(X)\}\right] = 0,$$

whatever be the choice of the function $f(\cdot)$ (i.e. regardless of whether $\mu_t^*(X)$ is modeled correctly). Therefore, we call (2.9) the *simple robust* (SR) influence function. We use (2.9) as a benchmark by which to gauge the efficiency of the "enriched" estimators of $\mu_{1t}^*$.

As for estimation, we estimate $\alpha_1^*$ from model $\mathcal{M}_\pi$ as the solution $\hat{\alpha}_1$ to

$$E_n\left[S_{\alpha_1}(T, R, X; \alpha_1)\right] = 0,$$

where

$$S_{\alpha_1}(T, R, X; \alpha_1) = R\left\{T - \pi_1(1; \alpha_1)\right\}.$$

In keeping with the "simple" aspect of this estimator, we estimate $\eta_t^*$ using data from the RCT only. Specifically, we can estimate $\eta_t^*$ using the estimating function

$$RI\{T=t\}S_{\eta_t}(Y,X;\eta_t)$$

This estimating function has mean zero at $\eta_t^*$ since assumptions (A1) – (A3) imply that

$$(R,T) \perp (Y_1, Y_0) | \, X.$$

Thus, we estimate $\eta_t^*$ as the solution $\hat{\eta}_t^S$ to

$$E_n\left[RI\{T=t\}S_{\eta_t}(Y,X;\eta_t)\right]=0.$$

Together, our simple robust estimator of $\mu_{1t}^*$ using only RCT data is the solution $\hat{\mu}_{1t}^{\text{RCT}-\text{SR}}$ to

$$E_n\left[\phi_{1t}^{\text{RCT}-\text{SR}}\{O;\mu_{1t},\hat{\rho}_t,\hat{\alpha}_1,\mu_t(X;\hat{\eta}_t^{\text{S}})\}\right]=0,$$

where $\hat{\rho}_t$ was defined previously.

Regardless of whether the working model for $\mu_t^*(X)$ is correctly specified, $\hat{\mu}_{1t}^{\text{RCT}-\text{SR}}$ is consistent and asymptotically normal, with the influence function

$$\begin{aligned}
&\text{IF}_{1t}^{\text{RCT-SR}}(O;\mu_{1t}^*,\rho_t^*,\alpha_1^*,\tilde{\eta}_t)\\
&=\phi_{1t}^{\text{RCT-SR}}\{O;\mu_{1t}^*,\rho_t^*,\alpha_1^*,\mu_t(X;\tilde{\eta}_t)\}\\
&\quad - E\left[\frac{\partial\phi_{1t}^{\text{RCT-SR}}\{O;\mu_{1t}^*,\rho_t^*,\alpha_1^*,\mu_t(X;\tilde{\eta}_t)\}}{\partial\alpha_1}\right]E\left[\frac{\partial S_{\alpha_1}(T,R,X;\alpha_1^*)}{\partial\alpha_1}\right]^{-1}S_{\alpha_1}(T,R,X;\alpha_1^*).
\end{aligned} \tag{2.10}$$

Note that the second term on the right hand side of equation (2.10) vanishes when $\tilde{\eta}_t = \eta_t^*$. As above, the robust asymptotic variance of $\hat{\mu}_{1t}^{\text{RCT}-\text{SR}}$ is $E[\text{IF}_{1t}^{\text{RCT-SR}}(O;\mu_{1t}^*,\rho_t^*,\alpha_1^*,\tilde{\eta}_t)^2]$, which can be estimated by

$E_n[\widehat{\mathrm{IF}}_{1t}^{\mathrm{RCT-SR}}(O;\hat{\mu}_{1t}^{\mathrm{RCT-SR}},\hat{\rho}_t,\hat{\alpha}_1,\hat{\eta}_t^{\mathrm{S}})^2]$, where $\widehat{\mathrm{IF}}_{1t}^{\mathrm{RCT-SR}}(O;\mu_{1t},\rho_t,\alpha_1,\eta_t)$ is the same as $\mathrm{IF}_{1t}^{\mathrm{RCT-SR}}(O;\mu_{1t},\rho_t,\alpha_1,\eta_t)$ except that the expectations are replaced by empirical averages.

### 2.3.3 Locally efficient estimators of $\mu_t^*$ and $\mu_{1t}^*$ under $\mathcal{M}_\pi \cap \mathcal{M}_\lambda$

The most efficient influence functions for estimating $\mu_t^*$ and $\mu_{1t}^*$ under the semiparametric model $\mathcal{P}_\pi$ were presented, respectively, in (2.5) and (2.7) above. Consistency of the associated estimators, $\hat{\mu}_t^{\mathrm{CC-LEDR}}$ and $\hat{\mu}_{1t}^{\mathrm{RCT-EDR}}$, requires correct specification of models for $\pi_t(0,X;\alpha_0^*)$ *and* $\lambda(X;\gamma^*)$, when the model for $\mu_t^*(X)$ is incorrectly specified.

It is natural to ask whether efficiency can be improved by further imposing modeling restriction $\mathcal{M}_\lambda$ on the observed data. That is, we consider the semiparametric model $\mathcal{P}_{\pi\lambda}$ which imposes restrictions on the distribution of the observed data through assumptions (A1) – (A3) and the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$). Let $\mathscr{T}_{\pi\lambda}^O$ denote the corresponding semiparametric tangent space. As discussed above, the most efficient influence function for a given parameter of interest is the projection (of the influence function of any RAL estimator of the parameter) onto $\mathscr{T}_{\pi\lambda}^O$. In Appendix II, we show

$$\mathscr{T}_{\pi\lambda}^{O,\perp} = \mathscr{T}_{\pi}^{O,\perp} \oplus \Lambda_c,$$

where $\mathscr{T}_{\pi}^{O,\perp}$ is defined as in (2.4) with $\lambda^*(X) = \lambda(X;\gamma^*)$, and

$$\Lambda_c = \left\{ \frac{R-\lambda(X;\gamma^*)}{\lambda(X;\gamma^*)\{1-\lambda(X;\gamma^*)\}} c(X) : E\left[\frac{\partial q(X;\gamma^*)}{\partial \gamma} c(X)\right] = 0 \right\}$$

To derive the most efficient influence functions for estimating $\mu_t^*$ and $\mu_{1t}^*$, we can project the influence functions for the CC-IPW and RCT-SIPW estimators

onto $\mathscr{T}^{O}_{\pi\lambda}$, respectively.

It can be shown that $\Pi\big[\mathrm{IF}_t^{\mathrm{CC-IPW}}(O;\mu_t^*,\alpha^*)\big|\,\Lambda_c\big]=0$. Thus, the most efficient influence function for estimating $\mu_t^*$ under $\mathcal{P}_{\pi\lambda}$ is equal to $\phi_t^{\mathrm{CC-LEDR}}\{O;\mu_t^*,\alpha^*,\lambda^*(X),\mu_t^*(X)\}$. Therefore, as mentioned in Section 2.3.2, the CC-LEDR estimator $\hat{\mu}_t^{\mathrm{CC-LEDR}}$ will be the locally efficient estimator of $\mu_t^*$ under the restrictions of assumption (A1) – (A3) and the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$), when both this joint model and the model for $\mu_t^*(X)$ are correctly specified.

In contrast, $\Pi\big[U_{1t}^{\mathrm{CC-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_c\big]\neq 0$. The most efficient influence

function for estimating $\mu_{1t}^*$ under $\mathcal{P}_{\pi\lambda}$ is the projection

$$
\begin{aligned}
&\phi_{1t}^{\mathrm{RCT-LE}}\{O;\mu_{1t}^*,\rho_t^*,\lambda^*,\alpha^*,\gamma^*,\mu_t^*(X)\} \\
&:= \Pi\big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\mathscr{T}_{\pi\lambda}^{O}\big] \\
&= U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*) - \Pi\Big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\mathscr{T}_{\pi\lambda}^{O,\perp}\Big] \\
&= U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*) - \Pi\big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\mathscr{T}_{\pi}^{O,\perp}\big] \\
&\quad - \Pi\big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_c\big] \\
&= \Pi\big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\mathscr{T}_{\pi}^{O}\big] - \Pi\big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_c\big] \\
&= \phi_{1t}^{\mathrm{RCT-EDR}}\{O;\mu_{1t}^*,\rho_t^*,\alpha^*,\lambda(X;\gamma^*),\mu_t^*(X)\} - \Pi\big[U_{1t}^{\mathrm{RCT-SIPW}}(O;\mu_{1t}^*,\rho_t^*)\big|\,\Lambda_c\big] \\
&= \frac{I\{T=t\}R}{\rho_t^*}(Y-\mu_{1t}^*) + (-1)^t R\left\{\frac{T-\pi_1(1;\alpha_1^*)}{\rho_t^*}\right\}\{\mu_t^*(X)-\mu_{1t}^*\} \\
&\quad - \left\{\frac{I\{T=t\}R}{\pi_t(1;\alpha_1^*)\lambda(X;\gamma^*)} - \frac{I\{T=t\}(1-R)}{\pi_t(0,X;\alpha_0^*)\{1-\lambda(X;\gamma^*)\}}\right\}\left\{\frac{Y-\mu_t^*(X)}{\rho_t^*}\right\} \\
&\qquad \times \left\{\frac{1}{\pi_t(1;\alpha_1^*)\lambda(X;\gamma^*)} + \frac{1}{\pi_t(0,X;\alpha_0^*)\{1-\lambda(X;\gamma^*)\}}\right\}^{-1} \\
&\quad - \{R-\lambda(X;\gamma^*)\}\left\{\frac{\mu_t^*(X)-\mu_{1t}^*}{\lambda^*} - k_t'\frac{\partial q(X;\gamma^*)}{\partial\gamma}\right\},
\end{aligned}
\tag{2.11}
$$

where

$$
\begin{aligned}
k_t' = E\left[\lambda(X;\gamma^*)\{1-\lambda(X;\gamma^*)\}\frac{\mu_t^*(X)-\mu_{1t}^*}{\lambda^*}\frac{\partial q(X;\gamma^*)}{\partial\gamma'}\right] \\
\times E\left[\lambda(X;\gamma^*)\{1-\lambda(X;\gamma^*)\}\frac{\partial q(X;\gamma^*)}{\partial\gamma}\frac{\partial q(X;\gamma^*)}{\partial\gamma'}\right]^{-1}.
\end{aligned}
$$

We call (2.11) the *locally efficient* (LE) influence function for estimating $\mu_{1t}^*$, which has mean 0 when evaluated at the truth $\{\mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, \mu_t^*(X)\}$. As for estimation, we still employ the parametric model $\mu_t^*(X) = \mu_t(X; \eta_t^*)$ and the corresponding estimator $\hat{\eta}_t$. Moreover, note that $\lambda^* = P[R = 1]$ can be directly estimated by the empirical average

$$\hat{\lambda} := E_n [I\{R = 1\}].$$

Then we define $\hat{\mu}_{1t}^{\text{RCT-LE}}$ as the solution to

$$E_n \left[ \phi_{1t}^{\text{RCT-LE}} \{O; \mu_{1t}, \hat{\rho}_t, \hat{\lambda}, \hat{\alpha}, \hat{\gamma}, \mu_t(X; \hat{\eta}_t)\} \right] = 0,$$

with the other estimators $\hat{\rho}_t, \hat{\alpha}, \hat{\gamma}$ defined previously.

It is important to notice that the influence function (2.11) is robust to misspecification of the model for $\mu_t^*(X)$ since

$$E \left[ \phi_{1t}^{\text{RCT-LE}} \{O; \mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, f(X)\} \right] = 0,$$

whatever be the choice of the function $f(\cdot)$. In other words, the resulting estimator will be consistent as long as the joint model $(\mathcal{M}_\pi \cap \mathcal{M}_\lambda)$ is correctly specified; and if the working model for $\mu_t^*(X)$ is correct as well, the resulting

estimator will be asymptotically efficient. As a result, under the correct specification of the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$), $\hat{\mu}_{1t}^{\text{RCT}-\text{LE}}$ has the influence function

$$\begin{aligned}
&\text{IF}_{1t}^{\text{RCT}-\text{LE}}(O; \mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, \tilde{\eta}_t) \\
&= \phi_{1t}^{\text{RCT}-\text{LE}}\{O; \mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, \mu_t(X; \tilde{\eta}_t)\} \\
&\quad - E\left[\frac{\partial \phi_{1t}^{\text{RCT}-\text{LE}}\{O; \mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, \mu_t(X; \tilde{\eta}_t)\}}{\partial \alpha'}\right] E\left[\frac{\partial S_\alpha(T, R, X; \alpha^*)}{\partial \alpha'}\right]^{-1} \\
&\qquad \times S_\alpha(T, R, X; \alpha^*) \\
&\quad - E\left[\frac{\partial \phi_{1t}^{\text{RCT}-\text{LE}}\{O; \mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, \mu_t(X; \tilde{\eta}_t)\}}{\partial \gamma'}\right] E\left[\frac{\partial S_\gamma(R, X; \gamma^*)}{\partial \gamma'}\right]^{-1} S_\gamma(R, X; \gamma^*)
\end{aligned} \tag{2.12}$$

where the last two terms on the right hand side of the equation vanish if $\tilde{\eta}_t = \eta_t^*$ (model for $\mu_t^*(X)$ is correct). We hereby obtain the "partially robust" asymptotic variance of $\hat{\mu}_{1t}^{\text{RCT}-\text{LE}}$ based on (2.12) as $E[\text{IF}_{1t}^{\text{RCT-LE}}(O; \mu_{1t}^*, \rho_t^*, \lambda^*, \alpha^*, \gamma^*, \tilde{\eta}_t)^2]$, since it is robust to misspecification of the model for $\mu_t^*(X)$ but still relies on correct specification of the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$); we estimate it by $E_n[\widehat{\text{IF}}_{1t}^{\text{RCT}-\text{LE}}(O; \hat{\mu}_{1t}^{\text{RCT}-\text{LE}}, \hat{\rho}_t, \hat{\lambda}, \hat{\alpha}, \hat{\gamma}, \tilde{\eta}_t)^2]$, where $\widehat{\text{IF}}_{1t}^{\text{RCT}-\text{LE}}(O; \mu_{1t}, \rho_t, \lambda, \alpha, \gamma, \eta_t)$ is the same as $\text{IF}_{1t}^{\text{RCT}-\text{LE}}(O; \mu_{1t}, \rho_t, \lambda, \alpha, \gamma, \eta_t)$ except that the expectations are replaced by empirical averages.

## 2.4 Simulation study

We present a simulation study to evaluate the finite sample performance of the six proposed estimators, i.e. CC-IPW and CC-LEDR for $\mu_t^*$ and RCT-SIW, RCT-SR, RCT-EDR and RCT-LE for $\mu_{1t}^*$. We also assess the robustness of our proposed estimation procedures by comparing the results under various types of model misspecification.

In this simulation, we generated 2000 datasets, each with a sample size $n = 500$. We considered the following true data generating mechanism:

- $X = (X_1, X_2)'$, $X_1 \sim$ Bernoulli(0.5), $X_2 \sim \mathcal{N}(0, 1)$ with $X_1$ independent of $X_2$
- Given $X$, the random variables $R$, $Y_1$ and $Y_0$ were assumed to be independent Bernoulli distributed with probabilities

$$P[R = 1|X] = \text{expit}(-0.5 + X_1 - X_2)$$

$$P[Y_1 = 1|X] = \text{expit}(-0.5 + X_1 + X_2)$$

$$P[Y_0 = 1|X] = \text{expit}(-1 + X_1 + 1.5X_2),$$

  respectively.

- Given $R, X, Y_1, Y_0$, the random variable $T$ was assumed to be Bernoulli $\{\pi_1^*(R, X)\}$, where

$$\pi_1^*(R, X) = R \cdot 0.5 + (1 - R) \cdot \text{expit}(X_1 + X_2).$$

- Set $Y = TY_1 + (1 - T)Y_0$.

Under these assumptions, $\lambda^* = 0.5$, $P[T = 1|R = 1] = 0.5$, $P[T = 1|R = 0] = 0.65$ and

$$\mu_1^* = 0.5, \qquad \mu_0^* = 0.4147, \qquad \Delta_{CC} = 0.0853;$$

$$\mu_{11}^* = 0.4415, \quad \mu_{10}^* = 0.3311, \quad \Delta_{RCT} = 0.1104.$$

In our simulations, we report average bias, average of standard error estimate, Monte Carlo standard deviation of the estimator, coverage of 95% Wald confidence intervals, Monte Carlo mean squared error of the estimator, and the relative efficiency with respect to the inverse probability weighted estimator. The relative efficiency (RE) was computed as the ratio of the mean squared errors (MSEs) between the CC-IPW (or RCT-SIPW) estimator and the estimator of interest for the same parameter. When applicable, we used estimates of asymptotic variances that are robust to model misspecification.

Table 2.2 shows that all estimators are unbiased under correct model specifications, for given sample size ($n = 500$). Moreover, the average of standard error estimates agrees well with the Monte Carlo standard deviation of the parameter estimators, and the coverage of the estimated 95% confidence intervals is close to their nominal level. As for the efficiency comparison, we observe from the MSE and RE columns that

- the CC-LEDR estimator is more efficient than the CC-IPW estimator
- the RCT-SR estimator improves the efficiency compared to the RCT-SIPW estimator
- the RCT-EDR and RCT-LE estimators yield large efficiency gains compared to the RCT-SIPW estimator – doubling the efficiency for estimating $\Delta_{RCT}$, which is equivalent (asymptotically) to a 50% reduction in the sample size required to achieve a desired power
- the advantage of the RCT-LE estimator over the RCT- EDR estimator is minimal.

In Table 2.3, we present results when the model for $\mu_t^*(X)$ is misspecified. Specifically, we incorrectly dropped the continuous component $X_2$ from the models for both $\mu_1^*(X)$ and $\mu_0^*(X)$. The two inverse probability weighted estimators, CC-IPW and CC-SIPW, are unchanged since they do not utilize estimators of $\mu_t^*(X)$. As expected, all the estimators are unbiased with some loss in efficiency as compared to the results shown in Table 2.2. The average of estimated standard errors agrees well with the Monte Carlo standard deviations. The coverage rates of the 95% confidence intervals are roughly accurate, which implies good performance of our robust variance estimators (when applicable). In addition, the summary statistics for the RCT-EDR and RCT-LE estimators are nearly identical in Table 2.3 (any differences are due to differences in the computation of the standard error estimator). This is due to the fact that the RCT-LE estimator reduces to the RCT-EDR estimator when the working model for $\mu^*(X)$ involves only a binary covariate that is shared with model $\mathcal{M}_\lambda$. In this case, the empirical average of the last term in the RCT-LE estimating function (2.11) can be shown to equal zero, with $\hat{\gamma}$ plugged in for $\gamma^*$.

In Table 2.4, we consider the case where the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$) is misspecified. We incorrectly dropped the continuous component $X_2$ from models $\mathcal{M}_\pi$ and $\mathcal{M}_\lambda$. Compared to the results under correct model specification in Table 2.2, only the RCT-SIPW and RCT-SR estimators remain exactly unchanged since they don't utilize $\pi_t^*(0, X)$ or $\lambda^*(X)$. As expected, the CC-LEDR, RCT-SIPW, RCT-SR and RCT-EDR estimators are consistent, with proper coverage of the 95% confidence intervals. Interestingly, RCT-EDR maintains its advantage in terms of efficiency over RCT-SIPW and RCT-SR under model misspecification, despite the fact that this is not guaranteed by theory. This

finding may simply be an artifact of our particular simulation study.

Table 2.2: Finite sample properties comparison under correct model specifications

| Parameter | Estimator | Bias | Mean SE | SD | 95% CI coverage (%) | MSE | RE† |
|---|---|---|---|---|---|---|---|
| $\mu_1^*$ | CC-IPW | 0.0005 | 0.0300 | 0.0303 | 95.0 | 0.0009 | 1.000 |
| | CC-LEDR | 0.0003 | 0.0286 | 0.0290 | 94.6 | 0.0008 | 1.088 |
| $\mu_0^*$ | CC-IPW | 0.0009 | 0.0380 | 0.0399 | 93.6 | 0.0016 | 1.000 |
| | CC-LEDR | 0.0012 | 0.0319 | 0.0317 | 95.2 | 0.0010 | 1.584 |
| $\Delta_{CC}$ | CC-IPW | -0.0004 | 0.0466 | 0.0484 | 94.0 | 0.0023 | 1.000 |
| | CC-LEDR | -0.0009 | 0.0400 | 0.0404 | 94.6 | 0.0016 | 1.433 |
| $\mu_{11}^*$ | RCT-SIPW | 0.0003 | 0.0443 | 0.0437 | 94.9 | 0.0019 | 1.000 |
| | RCT-SR | 0.0003 | 0.0420 | 0.0415 | 95.0 | 0.0017 | 1.104 |
| | RCT-EDR | 0.0001 | 0.0333 | 0.0336 | 95.0 | 0.0011 | 1.686 |
| | RCT-LE | 0.0001 | 0.0333 | 0.0336 | 94.6 | 0.0011 | 1.687 |
| $\mu_{10}^*$ | RCT-SIPW | 0.0009 | 0.0421 | 0.0412 | 95.7 | 0.0017 | 1.000 |
| | RCT-SR | 0.0008 | 0.0389 | 0.0381 | 95.5 | 0.0015 | 1.167 |
| | RCT-EDR | 0.0012 | 0.0328 | 0.0323 | 94.8 | 0.0010 | 1.629 |
| | RCT-LE | 0.0013 | 0.0326 | 0.0322 | 94.8 | 0.0010 | 1.633 |
| $\Delta_{RCT}$ | RCT-SIPW | -0.0006 | 0.0612 | 0.0598 | 95.0 | 0.0036 | 1.000 |
| | RCT-SR | -0.0006 | 0.0534 | 0.0524 | 95.3 | 0.0027 | 1.302 |
| | RCT-EDR | -0.0011 | 0.0420 | 0.0417 | 95.0 | 0.0017 | 2.051 |
| | RCT-LE | -0.0012 | 0.0418 | 0.0417 | 95.0 | 0.0017 | 2.056 |

† RE: relative efficiency with respect to the CC-IPW or RCT-SIPW estimator for the same parameter, computed as ratio of MSEs

## 2.5 Analysis of BARI

BARI (Bypass Angioplasty Revascularization Investigation) was designed to compare survival in patients receiving either percutaneous transluminal coronary angioplasty (PTCA) or coronary artery bypass grafting (CABG). As summarized in Brooks et al. (2000), a comprehensive cohort design was adopted to include 3,839 patients who were clinically eligible with severe angina or ischemia and multivessel coronary artery disease suitable for initial revascularization by either PTCA or CABG, and willing to be followed up. Among these patients, 1,829 patients consented to randomization and entered a randomized trial. The

Table 2.3: Finite sample properties comparison when $\mu_t^*(X)$ is misspecified

| Parameter | Estimator | Bias | Mean SE | SD | 95% CI coverage (%) | MSE | RE† |
|---|---|---|---|---|---|---|---|
| $\mu_1^*$ | CC-IPW | 0.0005 | 0.0300 | 0.0303 | 95.0 | 0.0009 | 1.000 |
| | CC-LEDR | 0.0005 | 0.0293 | 0.0296 | 95.0 | 0.0009 | 1.045 |
| $\mu_0^*$ | CC-IPW | 0.0009 | 0.0380 | 0.0399 | 93.6 | 0.0016 | 1.000 |
| | CC-LEDR | 0.0009 | 0.0356 | 0.0359 | 94.3 | 0.0013 | 1.233 |
| $\Delta_{CC}$ | CC-IPW | -0.0004 | 0.0466 | 0.0484 | 94.0 | 0.0023 | 1.000 |
| | CC-LEDR | -0.0004 | 0.0448 | 0.0456 | 94.8 | 0.0021 | 1.125 |
| $\mu_{11}^*$ | RCT-SIPW | 0.0003 | 0.0443 | 0.0437 | 94.9 | 0.0019 | 1.000 |
| | RCT-SR | 0.0002 | 0.0436 | 0.0429 | 95.1 | 0.0018 | 1.034 |
| | RCT-EDR | 0.0006 | 0.0345 | 0.0348 | 94.6 | 0.0012 | 1.577 |
| | RCT-LE | 0.0006 | 0.0345 | 0.0348 | 94.8 | 0.0012 | 1.577 |
| $\mu_{10}^*$ | RCT-SIPW | 0.0009 | 0.0421 | 0.0412 | 95.7 | 0.0017 | 1.000 |
| | RCT-SR | 0.0010 | 0.0416 | 0.0409 | 95.7 | 0.0017 | 1.012 |
| | RCT-EDR | 0.0013 | 0.0345 | 0.0344 | 94.6 | 0.0012 | 1.434 |
| | RCT-LE | 0.0013 | 0.0345 | 0.0344 | 94.5 | 0.0012 | 1.434 |
| $\Delta_{RCT}$ | RCT-SIPW | -0.0006 | 0.0612 | 0.0598 | 95.0 | 0.0036 | 1.000 |
| | RCT-SR | -0.0008 | 0.0595 | 0.0585 | 95.8 | 0.0034 | 1.045 |
| | RCT-EDR | -0.0007 | 0.0463 | 0.0465 | 94.6 | 0.0022 | 1.653 |
| | RCT-LE | -0.0007 | 0.0463 | 0.0465 | 94.5 | 0.0022 | 1.653 |

† RE: relative efficiency with respect to the CC-IPW or RCT-SIPW estimator for the same parameter, computed as ratio of MSEs

Table 2.4: Finite sample properties comparison
when the joint model ($\mathcal{M}_\pi \cap \mathcal{M}_\lambda$) is misspecified

| Parameter | Estimator | Bias | Mean SE | SD | 95% CI coverage (%) | MSE | RE[†] |
|---|---|---|---|---|---|---|---|
| $\mu_1^*$ | CC-IPW | 0.0261 | 0.0296 | 0.0300 | 85.7 | 0.0016 | 1.000 |
| | CC-LEDR | 0.0003 | 0.0285 | 0.0290 | 94.6 | 0.0008 | 1.879 |
| $\mu_0^*$ | CC-IPW | -0.0628 | 0.0346 | 0.0348 | 56.4 | 0.0051 | 1.000 |
| | CC-LEDR | 0.0011 | 0.0316 | 0.0314 | 95.2 | 0.0010 | 5.217 |
| $\Delta_{CC}$ | CC-IPW | 0.0889 | 0.0453 | 0.0460 | 51.0 | 0.0100 | 1.000 |
| | CC-LEDR | -0.0008 | 0.0398 | 0.0402 | 94.8 | 0.0016 | 6.207 |
| $\mu_{11}^*$ | RCT-SIPW | 0.0003 | 0.0443 | 0.0437 | 94.9 | 0.0019 | 1.000 |
| | RCT-SR | 0.0003 | 0.0420 | 0.0415 | 95.0 | 0.0017 | 1.104 |
| | RCT-EDR | 0.0001 | 0.0332 | 0.0334 | 95.0 | 0.0011 | 1.705 |
| | RCT-LE | 0.0795 | 0.0287 | 0.0297 | 22.0 | 0.0072 | 0.265 |
| $\mu_{10}^*$ | RCT-SIPW | 0.0009 | 0.0421 | 0.0412 | 95.7 | 0.0017 | 1.000 |
| | RCT-SR | 0.0008 | 0.0389 | 0.0381 | 95.5 | 0.0015 | 1.167 |
| | RCT-EDR | 0.0013 | 0.0327 | 0.0322 | 94.7 | 0.0010 | 1.630 |
| | RCT-LE | 0.1020 | 0.0324 | 0.0325 | 11.8 | 0.0115 | 0.148 |
| $\Delta_{RCT}$ | RCT-SIPW | -0.0006 | 0.0612 | 0.0598 | 95.0 | 0.0036 | 1.000 |
| | RCT-SR | -0.0006 | 0.0534 | 0.0524 | 95.3 | 0.0027 | 1.302 |
| | RCT-EDR | -0.0011 | 0.0419 | 0.0416 | 95.4 | 0.0017 | 2.066 |
| | RCT-LE | -0.0226 | 0.0397 | 0.0408 | 91.1 | 0.0022 | 1.641 |

[†] RE: relative efficiency with respect to the CC-IPW or RCT-SIW estimator for the same parameter, computed as ratio of MSEs

remaining 2,010 patients refused randomization but agreed to participate in the BARI registry, in which patients could choose their initial treatment in consultation with their physician. The follow-up plan was similar for randomized (RCT) and registry (OBS) patients.

In the OBS, "treatment" was defined as the first revascularization treatment received in the 3-month interval after study entry. One hundred and ninety-size registry patients did not receive any revascularization procedure during the 3-month interval and were excluded from the analysis. Also excluded were 33 RCT patients who did not receive their assigned treatment. While these exclusions are a strict departure from the intention-to-treat principle, they are commonly employed in treatment trials (often called modified intention-to-treat). Thus, a total of 3,610 patients were considered for inclusion in our analysis (1,814 from OBS and 1,796 from RCT). Among RCT patients, 904 (50.3%) were randomized to receive PTCA. Among OBS patients, 1,189 (65.5%) chose and received PTCA. These proportions are comparable to what we used in our simulation study.

The BARI Investigators (1996) reported 5-year mortality results for the RCT and compared the treatment groups using the logrank test. Feit et al. (2000) reported 7-year mortality results for the OBS and used Cox proportional hazards regression to computed adjusted (for baseline risk factors) differences between treatment groups. Neither paper found a statistically significant difference between PTCA and CABG in terms of mortality. For approximately 650 patients with treated diabetes, Detre et al. (1999) estimated unadjusted treatment effects among RCT patients and adjusted treatment effects (using multivariable Cox proportional hazards regression) among OBS patients. For RCT patients

(173 CABG, 170 PTCA), they found a significant increased risk of mortality for PTCA vs. CABG. For OBS patients (117 CABG, 182 PTCA), there was an increased risk but not statistically significant. Focusing on 5-year all-cause and cardiac mortality among all eligible and enrolled patients, Brooks et al. (2000) combined the RCT and OBS data to build a multivariable Cox proportional hazards regression model using randomization consent status, treatment group, baseline risk factors and two-way interactions. Their main goal was to identify risk factors for mortality as well as identify subgroup effects. They found that diabetic patients receiving insulin were at significantly increased risk for both all-cause and cardiac mortality if treated with PTCA vs. CABG. Further, patients with ST elevation were at increased risk of of cardiac mortality if treated with CABG vs. PTCA.

In our re-analysis, we aim to estimate the comprehensive cohort and the randomized trial causal effects of *initial* PTCA versus CABG treatment on 5-year mortality. The word "initial" is used to emphasize the first revascularization treatment (PTCA vs. CABG) received, as subsequent revascularization procedures (possibly different from the first treatment) actually occurred in some patients. Four censored patients had follow-up time less than 5 years and were excluded from our analysis.

In our analysis, $R$ denotes the randomization consent indicator, $T$ denotes the indicator of receiving PTCA, $Y$ denotes the binary indicator of death by the end of 5 years and $X$ is a vector of 17 baseline variables, including age, gender, race, highest level of education, number of diseased vessels, category of qualifying symptoms, levels of self reported health, systolic blood pressure, diastolic

blood pressure, and a set of indicators for proximal left anterior descending disease, prior myocardial infarction, heart failure, history of diabetes, history of treated diabetes, current smoking, hypertension, and renal dysfunction. Other baseline covariates with missingness greater than 10% for the whole study population were dropped from our analysis. For illustrative purposes only, subjects with missingness in any of the 17 baseline covariates were excluded, resulting in an analysis of 3,279 patients (90.8% of the sample after the initial exclusions), among which 1,518 were from OBS and 1,761 from RCT.

We fit logistic regression models for $\lambda^*(X)$ and $\pi_t^*(0, X)$. For the CC-LEDR, RCT-EDR and RCT-LE estimators, logistic regression models were fit for $\mu_1^*(X)$ and $\mu_0^*(X)$ based on all the data. For the RCT-SR estimator, these latter models were fit for $\mu_1^*(X)$ and $\mu_0^*(X)$ based only on the RCT data. For all the models, an intercept and main effect terms for $X$ were included.

Table 2.5 displays the estimated comprehensive cohort and randomized trial causal effects for the different estimation procedures. In addition to the point estimates and associated standard errors (SE), we also report 95% Wald-based confidence intervals (CI), two-sided p-values, and the relative efficiency (RE) with respect to the inverse probability weighted estimator of the same parameter. Here, RE is defined as the ratio of the estimated variance (squared standard error) between the CC-IPW (or RCT-SIPW) estimator and the estimator of interest for the same parameter.

For the comprehensive cohort causal effect, the CC-LEDR estimator is closer to the null and slightly more efficient than the CC-IPW estimator. Neither approach indicates a statistically significant difference between treatment groups with respect to five-year mortality for the entire cohort. For the randomized trial

causal effect, the RCT-EDR and RCT-LE is much closer to the null and much more efficient than the RCT-SIPW and RCT-SR estimators. In fact, inference based on the simple estimators are of borderline statistically significant, while those based on the enriched estimators are not. These results confirm that there are no statistically significant differences between treatment groups with respect to five-year mortality for the RCT cohort. Interestingly, the efficient comprehensive cohort (CC-LEDR) and randomized causal effect (RCT-EDR, RCT-LE) estimates are virtually identical, whereas the inefficient estimators were more disparate. In sum, our findings are consistent with the findings of the BARI Investigators (1996) and Feit et al. (2000).

Table 2.5: Comprehensive cohort and randomized trial causal effect of PTCA vs. CABG on 5-year mortality (%) for BARI

| Mortality (%) of interest | Estimator | Point estimate | SE | 95% CI | P-value | RE[†] |
|---|---|---|---|---|---|---|
| PTCA | CC-IPW | 11.0 | 0.7 | (9.5, 12.4) | <0.001 | 1.000 |
| (comprehensive) | CC-LEDR | 10.8 | 0.7 | (9.4, 12.2) | <0.001 | 1.087 |
| CABG | CC-IPW | 9.1 | 0.8 | (7.6, 10.6) | <0.001 | 1.000 |
| (comprehensive) | CC-LEDR | 9.4 | 0.8 | (8.0, 10.9) | <0.001 | 1.031 |
| PTCA-CABG | CC-IPW | 1.9 | 1.1 | (-0.2, 4.0) | 0.074 | 1.000 |
| (comprehensive) | CC-LEDR | 1.4 | 1.0 | (-0.6, 3.3) | 0.182 | 1.100 |
| PTCA | RCT-SIPW | 13.3 | 1.1 | (11.1, 15.5) | <0.001 | 1.000 |
| (randomized) | RCT-SR | 13.1 | 1.1 | (11.0, 15.3) | <0.001 | 1.098 |
| | RCT-EDR | 11.5 | 0.8 | (10.0, 13.1) | <0.001 | 2.071 |
| | RCT-LE | 11.5 | 0.8 | (10.0, 13.0) | <0.001 | 2.099 |
| CABG | RCT-SIPW | 10.3 | 1.0 | (8.3, 12.3) | <0.001 | 1.000 |
| (randomized) | RCT-SR | 10.4 | 1.0 | (8.4, 12.3) | <0.001 | 1.060 |
| | RCT-EDR | 10.1 | 0.8 | (8.5, 11.8) | <0.001 | 1.556 |
| | RCT-LE | 10.2 | 0.8 | (8.6, 11.8) | <0.001 | 1.571 |
| PTCA-CABG | RCT-SIPW | 3.0 | 1.5 | (-0.0, 6.0) | 0.050 | 1.000 |
| (randomized) | RCT-SR | 2.7 | 1.4 | (-0.0, 5.5) | 0.054 | 1.159 |
| | RCT-EDR | 1.4 | 1.1 | (-0.8, 3.5) | 0.205 | 1.988 |
| | RCT-LE | 1.3 | 1.1 | (-0.8, 3.4) | 0.226 | 2.004 |

[†] RE: relative efficiency with respect to the CC-IPW or RCT-SIPW estimator for the same parameter, computed as ratio of squared SEs.

## 2.6 Conclusion and Discussion

In this paper, we introduced new procedures for estimating causal effects of a binary treatment in the comprehensive cohort study (CCS) design. We introduced two estimators of the comprehensive cohort causal effect and four estimators for the randomized trial causal effect. Based on our simulation and data analysis, we recommend the CC-LEDR estimator for the comprehensive cohort causal effect, and the RCT-EDR estimator for the randomized trial causal effect. These estimators have improved efficiency relative to their inverse probability weighting counterparts and they offer robustness to model mis-specification. The RCT-SR estimator does provide efficiency and robustness over the RCT-SIPW estimator, but by using only RCT data it cannot compete with the RCT-EDR estimator. Compared to the RCT-EDR estimator, the advantage of the RCT-LE estimator in terms of efficiency appears limited, and it does not offer advantages in terms of robustness.

In specific regards to estimation of the randomized trial causal effect, there are important open questions that require further exploration. While it appears in our simulation study and data analysis that use of the RCT-EDR estimator results in a precision gain (relative to the RCT-SIPW estimator) comparable to the increase in observations from the OBS, this is not necessarily the case. In fact, there is no theoretical guarantee that the RCT-EDR will be be more precise than RCT-SIPW, in the presence of model misspecification. In future work, it will be useful to explore whether the ideas of Tan (2006), Tan (2010) and Rotnitzky et al. (2012) can be used to create an enriched doubly robust estimator that is guaranteed to be more efficient than RCT-SIPW. In addition,

it will be useful to develop a method for designing CCS's that properly balance statistical efficiency and logistics. We also note that Colantuoni and Rosenblum (2015) have proposed methods for leveraging prognostic baseline variables in randomized trials to construct more efficient estimators than RCT-SIPW. In future work, it will be useful to compare the RCT-EDR estimator to their estimator and see if it is possible to build an enriched doubly robust estimator that is guaranteed to have better efficiency properties.

Finally, for both estimands, the methods in this paper need to be extended to deal with time-to-event endpoints and missing data, both for outcomes and covariates.

## 2.7 Appendices

### 2.7.1 Appendix I: The space $\mathscr{T}_{\pi}^{O,\perp}$

To clarify the data structure, we first define the full data $F$ and the complete data $L$ as:

$$F := (X', Y_1, Y_0)'$$

$$L := (X', Y_1, Y_0, R, T)'$$

We rewrite the observed data $O$ in terms of the potential outcomes as:

$$O = (X', TY_1, (1-T)Y_0, R, T)'.$$

To study the semi-parametric tangent space for the observed data, we start with the restrictions on the complete data. Given assumptions (A1) – (A3), the likelihood of the complete data takes the form

$$f(X, Y_1, Y_0) \cdot P(R|X) \cdot P(T|R=1)^R \cdot P(T|R=0, X)^{1-R}.$$

Under the additional model restriction $\mathcal{M}_\pi$, we can express the semi-parametric tangent space for the complete data as:

$$\mathscr{T}_\pi^L = \mathscr{T}^F \oplus \mathscr{T}^{R|X} \oplus \mathscr{T}_\pi^{T|R,X},$$

where

$$\mathscr{T}^F = \{a(X, Y_1, Y_0) : E\left[a(X, Y_1, Y_0)\right] = 0\}$$

$$\mathscr{T}^{R|X} = \{a(X)\left\{R - \lambda^*(X)\right\} : \text{for any } a(X)\}$$

$$\mathscr{T}_\pi^{T|R,X} = \left\{k' \frac{\partial l(R, X; \alpha^*)}{\partial \alpha}\left\{T - \pi_1(R, X; \alpha^*)\right\} : \text{for all } k\right\}$$

Using Tsiatis (2006), the corresponding semi-parametric tangent space for the observed data can be shown to be

$$\mathscr{T}_\pi^O = \mathcal{T}^F \oplus \mathscr{T}^{R|X} \oplus \mathscr{T}_\pi^{T|R,X},$$

where

$$\mathcal{T}^F = \left\{E\left[a(X, Y_1, Y_0)|\, O\right] : \text{for all } a(X, Y_1, Y_0) \in \mathscr{T}^F\right\}.$$

We then derive

$$\mathscr{T}_\pi^{O,\perp} = \mathcal{T}^{F,\perp} \cap \mathscr{T}^{\{R|X\},\perp} \cap \mathscr{T}_\pi^{\{T|R,X\},\perp}.$$

It can be shown that

$$\begin{aligned}
\mathcal{T}^{F,\perp} &= \left\{h(O) : E\left[h(O)|\, F\right] \perp \mathscr{T}^F,\ E[h(O)] = 0\right\} \\
&= \{h(O) : E\left[h(O)|\, F\right] = 0\} \\
&= \left\{\sum_{\tau=0}^{1} \phi_{2\tau}\{O; \alpha^*, \lambda^*(X), h_\tau\} + \phi_3\{O; \alpha_0^*, \lambda^*(X), b\} : \text{for all } h_0, h_1, b\right\},
\end{aligned}$$

where

$$\phi_{2\tau}\{O;\alpha^*,\lambda^*(X),h_\tau\} = \left\{\frac{I\{T=\tau\}R}{\pi_\tau(1;\alpha_1^*)\lambda^*(X)} - \frac{I\{T=\tau\}(1-R)}{\pi_\tau(0,X;\alpha_0^*)\{1-\lambda^*(X)\}}\right\} h_\tau(X,Y_\tau)$$

$$\phi_3\{O;\alpha_0^*,\lambda^*(X),b\} = \frac{1-R}{1-\lambda^*(X)}\frac{T-\pi_1(0,X;\alpha_0^*)}{\pi_1(0,X;\alpha_0^*)\pi_0(0,X;\alpha_0^*)}b(X)$$

Moreover, in order for any $h(O) \in \mathcal{T}^{F,\perp}$ to be orthogonal to $\mathscr{T}^{R|X}$ and $\mathscr{T}_\pi^{T|R,X}$, $h_0, h_1, b$ must satisfy the following conditions:

$$E\left[h_0(X,Y_0) + h_1(X,Y_1)\middle|\, X\right] = 0$$

$$E\left[\frac{\partial l_0(X;\alpha_0^*)}{\partial\alpha_0}\left\{b(X) - \pi_0(0,X;\alpha_0^*)h_1(X,Y_1) + \pi_1(0,X;\alpha_0^*)h_0(X,Y_0)\right\}\right] = 0$$

$$E\left[\pi_0(1;\alpha_1^*)h_1(X,Y_1) - \pi_1(1;\alpha_1^*)h_0(X,Y_0)\right] = 0,$$

which are jointly equivalent to

$$E\left[h_1(X,Y_1)\middle|\, X\right] = -E\left[h_0(X,Y_0)\middle|\, X\right]$$

$$E\left[\frac{\partial l_0(X;\alpha_0^*)}{\partial\alpha_0}\left\{b(X) + E\left[h_0(X,Y_0)\middle|\, X\right]\right\}\right] = 0$$

$$E\left[h_1(X,Y_1)\right] = E\left[h_0(X,Y_0)\right] = 0$$

Writing $h_\tau(X,Y_\tau) = \{h_\tau(X,Y_\tau) - E\left[h_\tau(X,Y_\tau)\middle|\, X\right]\} + E\left[h_\tau(X,Y_\tau)\middle|\, X\right]$, we can express the orthogonal complement of the semi-parametric tangent space as

$$\mathscr{T}_\pi^{O,\perp} = \Lambda_h \oplus \Lambda_b \oplus \Lambda_a,$$

where

$$\Lambda_h = \Lambda_{h_0} \oplus \Lambda_{h_1}$$

$$\Lambda_{h_\tau} = \{\phi_{2\tau}\{O; \alpha^*, \lambda^*(X), h_\tau\} : E\left[h_\tau(X, Y_\tau)|\, X\right] = 0\} \quad \tau = 0, 1.$$

$$\Lambda_b = \left\{\phi_3\{O; \alpha_0^*, \lambda^*(X), b\} : E\left[\frac{\partial l_0(X; \alpha_0^*)}{\partial \alpha_0} b(X)\right] = 0\right\}$$

$$\Lambda_a = \{\phi_4\{O; \alpha_1^*, \lambda^*(X), a\} : E\left[a(X)\right] = 0\}$$

for

$$\phi_4\{O; \alpha_1^*, \lambda^*(X), a\} = \frac{R}{\lambda^*(X)} \left\{\frac{T - \pi_1(1; \alpha_1^*)}{\pi_1(1; \alpha_1^*)\pi_0(1; \alpha_1^*)}\right\} a(X).$$

It is important to note that $\Lambda_{h_0}, \Lambda_{h_1}, \Lambda_b, \Lambda_a$ are pairwise orthogonal to each other.

### 2.7.2 Appendix II: The space $\mathscr{T}_{\pi\lambda}^{O,\perp}$

Continuing with the results in Appendix I, we derive the orthogonal complement of the semiparametric tangent space that imposes the additional model restriction $\mathcal{M}_\lambda$, i.e., $\mathscr{T}_{\pi\lambda}^{O,\perp}$. Under $\mathcal{M}_\lambda$,

$$\lambda^*(X) = \lambda(X; \gamma^*) = \frac{\exp\left\{q(X; \gamma^*)\right\}}{1 + \exp\left\{q(X; \gamma^*)\right\}}, \tag{2.13}$$

$\mathscr{T}^{R|X}$ is replaced by

$$\mathscr{T}_\lambda^{R|X} = \left\{k' \frac{\partial q(X; \gamma^*)}{\partial \gamma} \left\{R - \lambda(X; \gamma^*)\right\} : \text{for all } k\right\}.$$

and

$$\mathscr{T}_{\pi\lambda}^{O,\perp} = \mathcal{T}^{F,\perp} \cap \mathscr{T}_\lambda^{\{R|X\},\perp} \cap \mathscr{T}_\pi^{\{T|R,X\},\perp}.$$

In order for any $h(O) \in \mathcal{T}^{F,\perp}$ to be orthogonal to $\mathscr{T}_{\lambda}^{R|X}$ and $\mathscr{T}_{\pi}^{T|R,X}$, $h_0, h_1, b$ must now satisfy the following conditions:

$$E\left[\frac{\partial q(X;\gamma^*)}{\partial \gamma}\left\{h_0(X,Y_0)+h_1(X,Y_1)\right\}\right]=0$$

$$E\left[\frac{\partial l_0(X;\alpha_0^*)}{\partial \alpha_0}\left\{b(X)-\pi_0(0,X;\alpha_0^*)h_1(X,Y_1)+\pi_1(0,X;\alpha_0^*)h_0(X,Y_0)\right\}\right]=0$$

$$E\left[\pi_0(1;\alpha_1^*)h_1(X,Y_1)-\pi_1(1;\alpha_1^*)h_0(X,Y_0)\right]=0$$

Writing

$$\begin{aligned}
h_\tau(X,Y_\tau) &= h_\tau(X,Y_\tau)-E\left[h_\tau(X,Y_\tau)|\,X\right]+E\left[h_\tau(X,Y_\tau)|\,X\right] \\
&\quad -\pi_{1-\tau}(R,X;\alpha^*)E\left[h_\tau(X,Y_\tau)|\,X\right]+\pi_\tau(R,X;\alpha^*)E\left[h_{1-\tau}(X,Y_{1-\tau})|\,X\right] \\
&\quad +\pi_{1-\tau}(R,X;\alpha^*)E\left[h_\tau(X,Y_\tau)|\,X\right]-\pi_\tau(R,X;\alpha^*)E\left[h_{1-\tau}(X,Y_{1-\tau})|\,X\right] \\
&= \left\{h_\tau(X,Y_\tau)-E\left[h_\tau(X,Y_\tau)|\,X\right]\right\} \\
&\quad +\pi_\tau(R,X;\alpha^*)E\left[h_0(X,Y_0)+h_1(X,Y_1)|\,X\right] \\
&\quad +\left\{\pi_{1-\tau}(R,X;\alpha^*)E\left[h_\tau(X,Y_\tau)|\,X\right]-\pi_\tau(R,X;\alpha^*)E\left[h_{1-\tau}(X,Y_{1-\tau})|\,X\right]\right\},
\end{aligned}$$

we can express the orthogonal complement of the semi-parametric tangent space as

$$\mathscr{T}_{\pi\lambda}^{O,\perp}=\Lambda_h\oplus\Lambda_b\oplus\Lambda_a\oplus\Lambda_c,$$

where $\Lambda_h, \Lambda_b, \Lambda_a$ are the same as defined in Appendix I with (2.13), and

$$\Lambda_c=\left\{\phi_5(O;\gamma^*,c): E\left[\frac{\partial q(X;\gamma^*)}{\partial \gamma}c(X)\right]=0\right\}$$

for

$$\phi_5(O;\gamma^*,c)=\left\{\frac{R-\lambda(X;\gamma^*)}{\lambda(X;\gamma^*)\left\{1-\lambda(X;\gamma^*)\right\}}\right\}c(X)$$

Again, note that $\Lambda_{h_0},\Lambda_{h_1},\Lambda_b,\Lambda_a,\Lambda_c$ are pairwise orthogonal to each other.

### 2.7.3 Appendix III: Projections

We list here the projections of an arbitrary function of the observed data (with mean zero and finite variance) onto $\Lambda_{h_\tau}(\tau=0,1),\Lambda_b,\Lambda_a$ and $\Lambda_c$, respectively.

Let $\mathcal{H}^O$ denote the observed data Hilbert space of all one dimensional, mean-zero measurable functions of $O$ with finite variance, equipped with the covariance inner product.

**Proposition 1.** *For any $h(O)\in\mathcal{H}^O$,*

$$\begin{aligned}\Pi[h(O)|\,\Lambda_{h_\tau}]=&\left\{\frac{I\{T=\tau\}R}{\pi_\tau(1;\alpha_1^*)\lambda^*(X)}-\frac{I\{T=\tau\}(1-R)}{\pi_\tau(0,X;\alpha_0^*)\left\{1-\lambda^*(X)\right\}}\right\}\\&\times\left\{\frac{1}{\pi_\tau(1;\alpha_1^*)\lambda^*(X)}+\frac{1}{\pi_\tau(0,X;\alpha_0^*)\left\{1-\lambda^*(X)\right\}}\right\}^{-1}\\&\times\left\{C_h(X,Y_\tau)-E\left[C_h(X,Y_\tau)|\,X\right]\right\},\end{aligned}$$

*where*

$$C_h(X,Y_\tau)=E\left[h(O)\left\{\frac{I\{T=\tau\}R}{\pi_\tau(1;\alpha_1^*)\lambda^*(X)}-\frac{I\{T=\tau\}(1-R)}{\pi_\tau(0,X;\alpha_0^*)\left\{1-\lambda^*(X)\right\}}\right\}\middle|\,X,Y_\tau\right]$$

*for $\tau=0,1$.*

**Proposition 2.** *For any $h(O)\in\mathcal{H}^O$,*

$$\begin{aligned}\Pi[h(O)|\,\Lambda_b]=&\left\{\frac{1-R}{1-\lambda^*(X)}\right\}\left\{\frac{T-\pi_1(0,X;\alpha_0^*)}{\pi_1(0,X;\alpha_0^*)\pi_0(0,X;\alpha_0^*)}\right\}B(X)^{-1}\\&\times\left\{D_h(X)-d_h'\frac{\partial l_0(X;\alpha_0^*)}{\partial\alpha_0}\right\},\end{aligned}$$

*where*

$$B(X) = \left\{ \frac{1}{1-\lambda^*(X)} \right\} \left\{ \frac{1}{\pi_1(0,X;\alpha_0^*)\pi_0(0,X;\alpha_0^*)} \right\}$$

$$D_h(X) = E\left[ h(O) \frac{1-R}{1-\lambda^*(X)} \frac{T-\pi_1(0,X;\alpha_0^*)}{\pi_1(0,X;\alpha_0^*)\pi_0(0,X;\alpha_0^*)} \middle| X \right]$$

$$d_h' = E\left[ B(X)^{-1} D_h(X) \frac{\partial l_0(X;\alpha_0^*)}{\partial \alpha_0'} \right] E\left[ B(X)^{-1} \frac{\partial l_0(X;\alpha_0^*)}{\partial \alpha_0} \frac{\partial l_0(X;\alpha_0^*)}{\partial \alpha_0'} \right]^{-1}.$$

**Proposition 3.** *For any* $h(O) \in \mathcal{H}^O$,

$$\Pi[h(O)|\,\Lambda_a] = \left\{ \frac{R}{\lambda^*(X)} \right\} \left\{ \frac{T-\pi_1(1;\alpha_1^*)}{\pi_1(1;\alpha_1^*)\pi_0(1;\alpha_1^*)} \right\} A(X)^{-1} \left\{ G_h(X) - g_h \right\},$$

*where*

$$A(X) = \left\{ \frac{1}{\lambda^*(X)} \right\} \left\{ \frac{1}{\pi_1(1;\alpha_1^*)\pi_0(1;\alpha_1^*)} \right\}$$

$$G_h(X) = E\left[ h(O) \frac{R}{\lambda^*(X)} \frac{T-\pi_1(1;\alpha_1^*)}{\pi_1(1;\alpha_1^*)\pi_0(1;\alpha_1^*)} \middle| X \right]$$

$$g_h = E\left[ A(X)^{-1} G_h(X) \right] E\left[ A(X)^{-1} \right]^{-1}.$$

**Proposition 4.** *For any* $h(O) \in \mathcal{H}^O$,

$$\Pi[h(O)|\,\Lambda_c] = \left\{ \frac{R-\lambda(X;\gamma^*)}{\lambda(X;\gamma^*)\left\{1-\lambda(X;\gamma^*)\right\}} \right\} C(X)^{-1} \left\{ J_h(X) - k_h' \frac{\partial q(X;\gamma^*)}{\partial \gamma} \right\},$$

*where*

$$C(X) = \frac{1}{\lambda(X;\gamma^*)\left\{1-\lambda(X;\gamma^*)\right\}}$$

$$J_h(X) = E\left[ h(O) \frac{R-\lambda(X;\gamma^*)}{\lambda(X;\gamma^*)\left\{1-\lambda(X;\gamma^*)\right\}} \middle| X \right]$$

$$k_h' = E\left[ C(X)^{-1} J_h(X) \frac{\partial q(X;\gamma^*)}{\partial \gamma'} \right] E\left[ C(X)^{-1} \frac{\partial q(X;\gamma^*)}{\partial \gamma} \frac{\partial q(X;\gamma^*)}{\partial \gamma'} \right]^{-1}.$$

# Chapter 3

# Optimal Outcome-Dependent Two-Phase Sampling

## 3.1 Introduction

When drawing inference about treatment effects in observational studies, it is important to adjust for potential confounding variables. Thus, it is important that these variables be collected. Unfortunately, certain variables, e.g., information from medical records, may be expensive to collect and, due to budgetary restrictions, cannot be ascertained on all subjects in the study sample. To address this issue, the outcome-dependent two-phase sampling design has been proposed (see, for example, White (1982) for case-control studies).

In the first phase, a simple random sample is drawn from the source population, with information obtained from all subjects in the study on treatment $T$, outcome $Y$ and covariates $S$ which are inexpensive to measure. In the second phase, subjects are classified into strata according to $(S, Y, T)$ and a random subsample, called a validation sample, is drawn from each stratum with stratum-specific sampling probabilities. Covariates $W$ (expensive to obtain) are measured on subjects in the validation sample. The two-phase sampling design was

first introduced by Neyman (1938) and discussed in Cochran (1963) (referred to as double sampling for stratification). With well-defined strata and wisely-chosen stratum-specific sampling probabilities, such designs can yield efficient parameter estimates given constraints on the validation sample size.

The issue of how to choose the strata to maximize the efficiency of the design has a long history, starting with Neyman (1934). The issue is complicated even when the stratification is based on a single univariate continuous variable (see, for example, Baillargeon and Rivest (2009)).

In this paper, we are interested in the optimal sampling scheme once the strata have been determined. Breslow and Cain (1988) and Breslow and Chatterjee (1999) have argued for a "balanced" design in which approximately equal numbers per stratum are selected at the second phase. Our idea is motivated by Reilly and Pepe (1995) and Gilbert et al. (2014), where the optimal sampling scheme within stratum for fixed (expected) phase two sample size or budget was determined by minimizing the asymptotic variance of a given estimator.

Given the strata, the optimal choice of the stratum-specific sampling probabilities depends on the exact statistical procedure used, specifically, the estimator for the parameter of interest. As pointed out in Wang et al. (2009), previous statistical methods for outcome-dependent two-phase sampling studies mainly focused on consistent and efficient estimators for the regression parameters of the distribution of the observed outcome $Y$ given treatment $T$ and covariates $S, W$; examples include Cosslett (1981, 1983), White (1982), Fears and Brown (1986), Breslow and Cain (1988), Carroll and Wand (1991), Pepe and Fleming (1991), Scott and Wild (1991, 1997), Schill et al. (1993), Robins et al. (1995), Breslow and Holubkov (1997a,b), Lawless et al. (1999), Breslow (2000), Breslow

et al. (2003), Chatterjee et al. (2003), Weaver and Zhou (2005), etc.

In this paper, we follow the methods developed in Wang et al. (2009) and focus on the estimators of the causal effect of a binary treatment on an outcome of interest; specifically, the difference in the marginal mean of the potential outcomes under two competing treatments. To assess the asymptotic variance of the estimators, we consider a class of influence functions considered by Wang et al. (2009). We specifically focus on influence functions associated with the simple inverse probability weighted and enriched doubly robust estimators. Given data from the first phase, our goal is to find the best stratum-specific sampling scheme that minimizes the variance of a given estimator, subject to the constraint of the expected validation sample size.

The variance depends on the conditional distribution of the expense covariates given first phase data. This variance cannot be estimated without second phase data. To address this issue, we propose an intermediate step to collect information on expensive covariates on a stratified sample of patients. We use these intermediate data to estimate the variance we seek to minimize and compute the optimal sampling probabilities that are used for determining the final validation sample. This approach contrasts with the approach of Gilbert et al. (2014), who propose to estimate the unknown conditional distribution using subject matter knowledge and separate pilot data.

Intuitively, our optimization procedure will produce a sampling allocation scheme similar in spirit to the *Neyman allocation* (Neyman, 1934), which sets the sampling fraction per stratum proportional to the stratum-specific standard deviation of outcomes, i.e. strata with larger variances are sampled more heavily to reduce the variance of the corresponding stratum mean. In our algorithm,

the "stratum-specific standard deviation" refers to a similar stratum-specific quantity that involves only the relevant expression identified from the influence function of the resulting estimator. Since our optimization procedure involves stricter constraints than the Neyman allocation procedure, no closed-form analytical solution is provided. Instead, the optimal stratum-specific sampling probabilities can be determined numerically using off-the-shelf optimization software.

The paper is organized as follows. In Section 3.2 we introduce our methodology. Specifically, we introduce the notation and data structure, review the estimation procedures of Wang et al. (2009), extend their procedures to allow for estimated, rather than known, sampling probabilities, and discuss the optimization algorithm for identifying the optimal sampling scheme. For simplicity of the formulae, Section 3.2 focuses on inference about the mean of the potential outcome under a single treatment, not the causal effect of interest. However, the ideas naturally generalize. In Section 3.3, we present results of a simulation study. In Section 3.4, we illustrate our methods using data from an observational study of critically ill patients in which some patients were treated with right heart catheterization (RHC). This dataset has been used previously to demonstrate methods for estimating the causal effect of RHC on 30-day survival. We use these data to build and evaluate hypothetical two-phase sampling designs based on different casual effect estimation procedures. The final section is devoted to a discussion.

## 3.2 Methods

Using the notation and framework in Wang et al. (2009), assume that we observe $n$ independent and identically distributed copies of $O = (S', Y, T, V, VW')'$, where $S$ is the vector of inexpensive covariates, $Y$ the outcome, $T$ the treatment indicator, $V$ the validation indicator and $W$ is the vector of expensive covariates. Note that $(S', Y, T)'$ is collected on all $n$ subjects in the first phase, based on which a validation sample is randomly drawn in the second phase; and the expensive covariates $W$ are measured only on the validation sample. For the validation sample, suppose that we aim to select no more than $n_1$ subjects among the treated ($T = 1$) and $n_0$ among the untreated ($T = 0$).

According to the study design, we denote $q(S, Y, T) = P[V = 1|S, Y, T]$, i.e. the probability of being selected as part of the validation sample. Define $q_t(S, Y) = q(S, Y, t)$, $t = 0, 1$. The degrees of freedom of $q(S, Y, T)$ depends on the dimension of the first phase data. When it is high-dimensional, we will use dimension reduction/categorization procedures to classify subjects into a limited number ($K$) of strata. Let $r : (S, Y, T) \mapsto \{1, \ldots, K\}$ be the stratification function. For all subjects in stratum $k$ (i.e., $r(S, Y, T) = k$), we assume the probability of being selected into the validation sample is $q_k := P[V = 1|r(S, Y, T) = k]$. So, $q(S, Y, T) = \sum_{k=1}^{K} q_k I_{\{r(S,Y,T)=k\}}$.

Let $Y_1$ and $Y_0$ be the potential outcomes for treatment 1 and 0 respectively. We make the "stable unit treatment value" assumption (Rubin (1986)) and consistency assumption (i.e., $Y = TY_1 + (1 - T)Y_0$). The target parameter of interest is $\mu_t^* = E[Y_t]$. To identify $\mu_t^*$ from the observed data, we assume that $T$ is independent of $(Y_1, Y_0)$ given the complete set of covariates $X = (S', W')'$.

We denote the treatment assignment probability by $\pi_t^*(X) = P[T = t|X]$ and assume $\pi_t^*(x) > 0$ for all $x$ and $t = 0, 1$. Furthermore, we denote $\mu_t^*(X) = E[Y_t|X]$.

### 3.2.1 Estimators of $\mu_t^*$

Wang et al. (2009) developed regular and asymptotically linear (RAL) estimators of $\mu_t^*$ which depend on known $q(S, Y, T)$. The class of estimators considered by Wang et al. (2009) require specification of parametric models for $\pi_t^*(X) = \pi_t(X; \alpha^*)$ (treatment regression model) and $\mu_t^*(X) = \mu_t(X; \eta^*)$ (outcome regression model), where $\alpha^*$ and $\eta^*$ denote the true value of model parameters $\alpha$ and $\eta$, respectively. The influence functions for RAL estimators of $\mu_t^*$ take the form:

$$\mathrm{IF}_t\left(O; \mu_t^*, \tilde{\alpha}, \tilde{\eta}\right) = \frac{V}{q_T(S, Y)} h_{1t}\left(X, Y, T; \mu_t^*, \tilde{\alpha}, \tilde{\eta}\right) + h_{2t}\left(S, Y, T; \mu_t^*, \tilde{\alpha}, \tilde{\eta}\right), \quad (3.1)$$

where either $\tilde{\alpha} = \alpha^*$ (the treatment regression model is correct) or $\tilde{\eta} = \eta^*$ (the outcome regression model is correct); and $h_{1t}(\cdot), h_{2t}(\cdot)$ are specific functions not involving $q(S, Y, T)$. The associated asymptotic variance for an estimator of $\mu_t^*$ with influence function (3.1) is

$$\begin{aligned}
&E\left[\mathrm{IF}_t\left(O; \mu_t^*, \tilde{\alpha}, \tilde{\eta}\right)^2\right] \\
&= E\left[\frac{1}{q(S, Y, T)} E\left[h_{1t}(X, Y, T; \mu_t^*, \tilde{\alpha}, \tilde{\eta})^2 \middle| S, Y, T\right]\right] + C_t \\
&= \sum_{k=1}^{K} \frac{1}{q_k} E\left[h_{1t}(X, Y, T; \mu_t^*, \tilde{\alpha}, \tilde{\eta})^2 \,\middle|\, r(S, Y, T) = k\right] P\left[r(S, Y, T) = k\right] + C_t
\end{aligned}$$

for some constant $C_t$ that does not depend on $q(S, Y, T)$. Thus, minimizing the

asymptotic variance over $q(S,Y,T)$ is equivalent to minimizing

$$\sum_{k=1}^{K} \frac{1}{q_k} E\left[h_{1t}(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta})^2 \,\middle|\, r(S,Y,T)=k\right] P\left[r(S,Y,T)=k\right] \tag{3.2}$$

over $q_1,\ldots,q_K$.

We highlight two estimators: simple inverse probability weighted (SIPW) estimator and enriched doubly robust (EDR) estimator.

**SIPW estimator**

Wang et al. (2009) introduced the following *simple inverse probability weighted* estimating function:

$$U_t^{\text{SIPW}}(O;\mu_t,\pi_t) = \frac{VI_{\{T=t\}}}{q_t(S,Y)\pi_t(X)}(Y-\mu_t),$$

which has mean 0 evaluated when at the truth $\mu_t^*$ and $\pi_t^*(X)$. To estimate $\mu_t^*$ using this estimating function, Wang et al. (2009) posited a logistic regression model for $\pi_t^*(X)$ of the form:

$$\pi_t^*(X) = \pi_t(X;\alpha^*) = \frac{\exp\{tl(X;\alpha^*)\}}{1+\exp\{l(X;\alpha^*)\}},$$

where $l(X;\alpha)$ is a specified function of $X$ and $\alpha$ and $\alpha^*$ is the true value of $\alpha$. Then $(\mu_t^*,\alpha^*)$ is estimated by solving

$$E_n\left[\begin{array}{c} U_t^{\text{SIPW}}(O;\mu_t,\alpha) \\ \frac{V}{q_T(S,Y)} S_\alpha(T,X;\alpha) \end{array}\right] = 0, \tag{3.3}$$

where $E_n[\cdot]$ is the empirical expectation operator and

$$U_t^{\text{SIPW}}(O;\mu_t,\alpha) = \frac{VI_{\{T=t\}}}{q_t(S,Y)\pi_t(X;\alpha)}(Y-\mu_t)$$

$$S_\alpha(T,X;\alpha) = \frac{\partial l(X;\alpha)}{\partial \alpha}\left\{T-\pi_1(X;\alpha)\right\}$$

Let $(\widehat{\mu_t^{\text{SIPW}}}, \hat{\alpha})$ denote the solution to (3.3). As pointed out in Wang et al. (2009), $\widehat{\mu_t^{\text{SIPW}}}$ is a *regular and asymptotically linear* (RAL) estimator of $\mu_t^*$ with influence function

$$\text{IF}_t^{\text{SIPW}}(O; \mu_t^*, \alpha^*) = U_t{}^{\text{SIPW}}(O; \mu_t^*, \alpha^*) - \text{A}_t^{\text{SIPW}}(\mu_t^*, \alpha^*) \frac{V}{q_T(S,Y)} S_\alpha(T, X; \alpha^*)$$

where

$$\begin{aligned}
\text{A}_t^{\text{SIPW}}(\mu_t^*, \alpha^*) &:= E\left[\frac{\partial U_t{}^{\text{SIPW}}(O; \mu_t^*, \alpha^*)}{\partial \alpha'}\right] E\left[\frac{V}{q_T(S,Y)} \frac{\partial S_\alpha(T, X; \alpha^*)}{\partial \alpha'}\right]^{-1} \\
&= E\left[\frac{\partial}{\partial \alpha'}\left(\frac{I_{\{T=t\}}}{\pi_t(X; \alpha^*)}(Y - \mu_t^*)\right)\right] E\left[\frac{\partial S_\alpha(T, X; \alpha^*)}{\partial \alpha'}\right]^{-1}.
\end{aligned}$$

It is useful to note that $\text{A}_t^{\text{SIPW}}(\mu_t^*, \alpha^*)$ does not depend on $q(S, Y, T)$. Now, we can re-express $\text{IF}_t^{\text{SIPW}}(O; \mu_t^*, \alpha^*)$ as

$$\text{IF}_t^{\text{SIPW}}(O; \mu_t^*, \alpha^*) = \frac{V}{q_T(S,Y)} \left\{ \frac{I_{\{T=t\}}}{\pi_t(X; \alpha^*)}(Y - \mu_t^*) - \text{A}_t^{\text{SIPW}}(\mu_t^*, \alpha^*) S_\alpha(T, X; \alpha^*) \right\}$$

which is exactly in the form of (3.1), with

$$\begin{aligned}
h_{1t}(X, Y, T; \mu_t, \alpha, \eta) &= h_{1t}^{\text{SIPW}}(X, Y, T; \mu_t, \alpha) \\
&= \frac{I_{\{T=t\}}}{\pi_t(X; \alpha)}(Y - \mu_t) - \text{A}_t^{\text{SIPW}}(\mu_t, \alpha) S_\alpha(T, X; \alpha) \qquad (3.4) \\
h_{2t}(S, Y, T; \mu_t, \alpha, \eta) &= h_{2t}^{\text{SIPW}}(S, Y, T; \mu_t, \alpha) = 0
\end{aligned}$$

**EDR estimator**

Wang et al. (2009) proposed and recommended the *enriched doubly robust* estimator, which incorporates the information from the non-validation sample. In simulation studies, they showed that the EDR estimator is more efficient and robust to model misspecification than the SIPW estimator. In the construction

of their EDR estimator, Wang et al. (2009) posited, in addition to the model for $\pi_t^*(X)$, a model for $\mu_t^*(X) = \mu_t(X;\eta^*)$, where $\eta^*$ denotes the true value of the model parameter $\eta$. They introduced the following EDR estimating function:

$$U_t^{\mathrm{EDR}}(O;\mu_t,\alpha,\eta) = \frac{V}{q_T(S,Y)}\varphi_t(X,Y,T;\mu_t,\alpha,\eta) - \left\{\frac{V}{q_T(S,Y)} - 1\right\} E\left[\varphi_t(X,Y,T;\mu_t,\alpha,\eta)\middle| S,Y,T\right],$$

where

$$\varphi_t(X,Y,T;\mu_t,\alpha,\eta) = \frac{I_{\{T=t\}}}{\pi_t(X;\alpha)}(Y-\mu_t) + (-1)^t \left\{\frac{T-\pi_1(X;\alpha)}{\pi_t(X;\alpha)}\right\}\{\mu_t(X;\eta)-\mu_t\}.$$

They show that this estimating function was doubly robust in the sense that

$$E\left[U_t^{\mathrm{EDR}}(O;\mu^*,\tilde{\alpha},\tilde{\eta})\right] = 0,$$

when either $\tilde{\alpha} = \alpha^*$ (model for $\pi_t^*(X)$ is correctly specified) or $\tilde{\eta} = \eta^*$ (model $\mu_t^*(X)$ is correctly specified).

The parameters $(\mu_t^*, \alpha^*, \eta^*)$ are estimated by solving

$$E_n\begin{bmatrix} \widehat{U_t^{\mathrm{EDR}}}(O;\mu_t,\alpha,\eta) \\ \frac{V}{q_T(S,Y)} S_\alpha(T,X;\alpha) \\ \frac{VI_{\{T=t\}}}{q_T(S,Y)} S_{t\eta}(Y,X;\eta) \end{bmatrix} = 0, \tag{3.5}$$

where $S_{t\eta}(Y,X;\eta)$ is the corresponding score function for $\eta$ from the model for $\mu_t^*(X)$, $\widehat{U_t^{\mathrm{EDR}}}(O;\mu_t,\alpha,\eta)$ is the same as $U_t^{\mathrm{EDR}}(O;\mu_t,\alpha,\eta)$ except that $E\left[\varphi_t(X,Y,T;\mu_t,\alpha,\eta)\middle| S,Y,T\right]$ is replaced by an estimator, e.g.,

$$\sum_{k=1}^{K} I(r(S,Y,T)=k)\hat{E}\left[\frac{V}{q_T(S,Y)}\varphi_t(X,Y,T;\mu_t,\alpha,\eta)\middle| r(S,Y,T)=k\right].$$

Let $(\widehat{\mu_t^{\mathrm{EDR}}}, \hat{\alpha}, \hat{\eta})$ denote the solution to (3.5). In fact, when either $\tilde{\alpha} = \alpha^*$ or $\tilde{\eta} = \eta^*$, the estimator $\widehat{\mu_t^{\mathrm{EDR}}}$ will be consistent and asymptotically normal even if the estimator for $E\left[\varphi_t(X,Y,T;\mu_t,\alpha,\eta)\middle| S,Y,T\right]$ is biased.

The influence function for $\widehat{\mu_t^{\text{EDR}}}$ can be shown to be

$$\begin{aligned}
&\text{IF}_t^{\text{EDR}}(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}) \\
&= U_t^{\text{EDR}}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) - \text{A}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta})\frac{V}{q_T(S,Y)}S_\alpha(T,X;\tilde{\alpha}) \\
&\quad - \text{H}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta})\frac{VI_{\{T=t\}}}{q_T(S,Y)}S_{t\eta}(Y,X;\tilde{\eta})
\end{aligned}$$

where either $\tilde{\alpha} = \alpha^*$ or $\tilde{\eta} = \eta^*$, and

$$\begin{aligned}
\text{A}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta}) &:= E\left[\frac{\partial U_t^{\text{EDR}}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)}{\partial\alpha'}\right]E\left[\frac{V}{q_T(S,Y)}\frac{\partial S_\alpha(T,X;\tilde{\alpha})}{\partial\alpha'}\right]^{-1} \\
&= E\left[\frac{\partial\varphi_t\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)}{\partial\alpha'}\right]E\left[\frac{\partial S_\alpha(T,X;\tilde{\alpha})}{\partial\alpha'}\right]^{-1} \\
\text{H}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta}) &:= E\left[\frac{\partial U_t^{\text{EDR}}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)}{\partial\eta'}\right]E\left[\frac{VI_{\{T=t\}}}{q_T(S,Y)}\frac{\partial S_{t\eta}(Y,X;\tilde{\eta})}{\partial\eta'}\right]^{-1} \\
&= E\left[\frac{\partial\varphi_t\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)}{\partial\eta'}\right]E\left[I_{\{T=t\}}\frac{\partial S_{t\eta}(Y,X;\tilde{\eta})}{\partial\eta'}\right]^{-1}
\end{aligned}$$

Importantly, $\text{A}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta})$ and $\text{H}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta})$ do not depend upon $q(S,Y,T)$. We can re-express $\text{IF}_t^{\text{EDR}}(O;\mu_t^*,\tilde{\alpha},\tilde{\eta})$ as

$$\begin{aligned}
&\text{IF}_t^{\text{EDR}}(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}) \\
&= \frac{V}{q_T(S,Y)}\Big\{\varphi_t\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) - E\left[\varphi_t\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)\middle|\, S,Y,T\right] \\
&\qquad - \text{A}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta})S_\alpha(T,X;\tilde{\alpha}) - \text{H}_t^{\text{EDR}}(\mu_t^*,\tilde{\alpha},\tilde{\eta})I_{\{T=t\}}S_{t\eta}(Y,X;\tilde{\eta})\Big\} \\
&\quad + E\left[\varphi_t\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)\middle|\, S,Y,T\right],
\end{aligned}$$

which is also in the form of (3.1), with

$$\begin{aligned} h_{1t}(X,Y,T;\mu_t,\alpha,\eta) &= h_{1t}^{\mathrm{EDR}}(X,Y,T;\mu_t,\alpha,\eta) \\ &= \varphi_t\left(X,Y,T;\mu_t,\alpha,\eta\right) - E\left[\varphi_t\left(X,Y,T;\mu_t,\alpha,\eta\right)\middle|\, S,Y,T\right] \\ &\quad - \mathrm{A}_t^{\mathrm{EDR}}(\mu_t,\alpha,\eta)S_\alpha(T,X;\alpha) - \mathrm{H}_t^{\mathrm{EDR}}(\mu_t,\alpha,\eta)I_{\{T=t\}}S_{t\eta}(Y,X;\eta) \end{aligned} \tag{3.6}$$

$$\begin{aligned} h_{2t}(S,Y,T;\mu_t,\alpha,\eta) &= h_{2t}^{\mathrm{EDR}}(S,Y,T;\mu_t,\alpha,\eta) \\ &= E\left[\varphi_t\left(X,Y,T;\mu_t,\alpha,\eta\right)\middle|\, S,Y,T\right] \end{aligned}$$

**Empirical estimation of** $q(S,Y,T)$

The asymptotic variance of influence function (3.1) that was presented in (3.2) was computed under the assumption that $q(S,Y,T)$ is known and not estimated from the observed data. However, it can be shown that extra efficiency can be gained by considering $q(S,Y,T)$ to be unknown and using estimates in the solving (3.3) and (3.5) above.

Under our dimension reduction/categorization assumption, we know that $q(S,Y,T)$ is determined by the $q_k$'s, where $q_k = P[V=1|r(S,T,Y)=k]$. As a result, an unbiased estimating function for $q_k$ is

$$U_k(S,T,Y;q_k) = I_{\{r(S,T,Y)=k\}}(V-q_k)$$

and $q_k$ can be estimated as the solution $\hat{q}_k$ to

$$E_n[U_k(S,T,Y;q_k)] = 0$$

The influence function for an estimator based on (3.1) with $q(S,Y,T)$ are

replaced by $\sum_{k=1}^{K} \hat{q}_k I_{\{r(S,T,Y)=k\}}$ can be shown to have the form:

$$\begin{aligned}
\mathrm{IF}_t^{\dagger}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) &= \mathrm{IF}_t\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) \\
&\quad - \sum_{k=1}^{K} E\left[\frac{\partial \mathrm{IF}_t\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)}{\partial q_k}\right] E\left[\frac{\partial U_k(S,T,Y;q_k)}{\partial q_k}\right]^{-1} U_k(S,T,Y;q_k) \\
&= \frac{V}{q(S,Y,T)} h_{1t}\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) + h_{2t}\left(S,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) \\
&\quad - \left\{\frac{V}{q(S,Y,T)} - 1\right\} E\left[h_{1t}\left(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right) \middle| r(S,Y,T)\right],
\end{aligned} \tag{3.7}$$

where either $\tilde{\alpha} = \alpha^*$ or $\tilde{\eta} = \eta^*$.

Note that $\mathrm{IF}_t^{\dagger}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)$ is exactly $\mathrm{IF}_{\mathrm{t}}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)$ minus its projection onto the space spanned by $\{U_k(S,T,Y;q_k) : k = 1,\ldots,K\}$. By the triangle inequality, it can be shown that $\mathrm{IF}_t^{\dagger}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)$ will have a smaller variance than $\mathrm{IF}_{\mathrm{t}}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)$. The associated asymptotic variance of an estimator of $\mu_t^*$ with influence function (3.7) is

$$\begin{aligned}
&E\left[\mathrm{IF}_t^{\dagger}\left(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}\right)^2\right] \\
&= E\left[\frac{1}{q(S,Y,T)}\mathrm{Var}\left[h_{1t}(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}) \middle| r(S,Y,T)\right]\right] + C_t^{\dagger} \\
&= \sum_{k=1}^{K} \frac{1}{q_k}\mathrm{Var}\left[h_{1t}(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}) \,|\, r(S,Y,T)=k\right] P\left[r(S,Y,T)=k\right] + C_t^{\dagger}
\end{aligned}$$

for some constant $C_t^{\dagger}$ that does not depend on $q(S,Y,T)$. Thus, minimizing the asymptotic variance over $q(S,Y,T)$ is equivalent to minimizing

$$\sum_{k=1}^{K} \frac{1}{q_k}\mathrm{Var}\left[h_{1t}(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta}) \,|\, r(S,Y,T)=k\right] P\left[r(S,Y,T)=k\right] \tag{3.8}$$

over $q_1,\ldots,q_K$.

### 3.2.2 Optimizing the choice of $q_k$

We are interested in finding the optimal $q(S, Y, T)$ to draw inference about $\mu_t^* = E[Y_t]$. For a given estimator, our goal to find the choice of $q(S, Y, T)$ that minimizes the variance of the estimator subject to sample size restrictions on the validation sample. Formally, our goal is minimize either (3.2) for estimators which treat $q(S, T, Y)$ as known or (3.8) for estimators which treat $q(S, T, Y)$ as unknown with respect to $q(S, Y, T)$, provided that the expected validation sample size in treatment $t$ is less than $n_t$ for $t = 0, 1$. The expected validation sample size in treatment group $t$ can be expressed as:

$$nE\left[I_{\{T=t\}}q(S, Y, T)\right] = n\sum_{k=1}^{K} q_k P[T = t, r(S, Y, T) = k].$$

The optimization problem is

$$\begin{cases} \underset{q_1,\ldots,q_K}{\arg\min}(3.2) \left(\underset{q_1,\ldots,q_K}{\arg\min}(3.8)\right) \\ \text{subject to} \quad n\sum\limits_{k=1}^{K} q_k P\left[T = 1, r(S, Y, T) = k\right] \leq n_1 \\ \qquad\qquad\quad n\sum\limits_{k=1}^{K} q_k P\left[T = 0, r(S, Y, T) = k\right] \leq n_0 \\ \qquad\qquad\quad 0 \leq q_1, \ldots, q_K \leq 1 \end{cases} \tag{3.9}$$

To execution the optimization, the conditional expectation in (3.2) (or conditional variance in (3.8)) and probabilities $P\left[r(S, Y, T) = k\right]$, $P\left[T = 1, r(S, Y, T) = k\right]$ and $P\left[T = 0, r(S, Y, T) = k\right]$ need to be estimated. The probabilities can be estimated, using first stage data, by $E_n[I(r(S, Y, T) = k)]$, $E_n[I(T = 1, r(S, Y, T) = k)]$ $E_n[I(T = 0, r(S, Y, T) = k)]$ respectively. We discuss estimation of the conditional expectation in the next subsection.

**Intermediate step**

The conditional expectation $E\left[h_{1t}(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta})^2\middle|\, r(S,Y,T)=k\right]$ in (3.2) and conditional variance $\mathrm{Var}\left[h_{1t}(X,Y,T;\mu_t^*,\tilde{\alpha},\tilde{\eta})\,|\,r(S,Y,T)=k\right]$ in (3.8) depends on the conditional distribution of $W$ given $r(S,Y,T)=k$. This conditional distribution cannot be estimated based on the first phase data. To address this problem, we propose an intermediate step between the first and second phases, where we collect data on the expensive covariates $W$ from a small subset of subjects in first phase. With this intermediate step, we would have $n$ independent and identically distributed copies of $O^{(1)}=(S',Y,T,V^{(1)},V^{(1)}W')'$, where $V^{(1)}$ is the validation indicator in the intermediate step, i.e. the indicator of being selected into the small subset on which we collect $W$. We assume the same stratification function is used. For all subjects in stratum $k$ (i.e., $r(S,Y,T)=k$), we assume the pre-specified probability of being selected into the intermediate validation sample is $q_k^{(1)}:=P[V^{(1)}=1|r(S,Y,T)=k]$. We define $q^{(1)}(S,Y,T)=\sum_{k=1}^{K}q_k^{(1)}I_{\{r(S,Y,T)=k\}}$.

The advantage of the intermediate step is that the resulting observed data $O^{(1)}$ are generated by an outcome-dependent two-phase sampling scheme and we can use the techniques of Wang et al. (2009) to estimate $\mu_t^*$, $\alpha^*$ and $\eta^*$. Then, the conditional expectation in (3.2) can be estimated by

$$E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\hat{h}_{1t}\Big(X,Y,T;\hat{\mu}_t^{(1)},\hat{\alpha}^{(1)},\hat{\eta}^{(1)}\Big)^2\middle|\, r(S,Y,T)=k\right], \tag{3.10}$$

and the conditional variance in (3.8) can be estimated by

$$E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\hat{h}_{1t}\Big(X,Y,T;\hat{\mu}_t^{(1)},\hat{\alpha}^{(1)},\hat{\eta}^{(1)}\Big)^2\,\Big|\,r(S,Y,T)=k\right]$$

$$-E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\hat{h}_{1t}\Big(X,Y,T;\hat{\mu}_t^{(1)},\hat{\alpha}^{(1)},\hat{\eta}^{(1)}\Big)\,\Big|\,r(S,Y,T)=k\right]^2 \qquad (3.11)$$

where $E_n[\cdot|r(S,Y,T)=k]$ is the empirical expectation operator over the stratum $r(S,Y,T)=k$, $(\hat{\mu}_t^{(1)},\hat{\alpha}^{(1)},\hat{\eta}^{(1)})$ are estimators of $(\mu_t^*,\alpha^*,\eta^*)$ of based on $O^{(1)}$, and $\hat{h}_{1t}$ is an estimator of $h_{1t}$.

For the SIPW estimator, $h_{1t}$ is given by (3.4) and it can be estimated by replacing $\mathrm{A}_t^{\mathrm{SIPW}}(\mu_t,\alpha)$ with

$$\widehat{\mathrm{A}_t^{\mathrm{SIPW}}}(\mu_t,\alpha)=E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\frac{\partial}{\partial\alpha'}\left(\frac{I_{\{T=t\}}}{\pi_t(X;\alpha)}(Y-\mu_t)\right)\right]$$

$$\times E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\frac{\partial S_\alpha(T,X;\alpha^*)}{\partial\alpha'}\right]^{-1}.$$

For the EDR estimator, $h_{1t}$ is given by (3.6) and it can be estimated by replacing $\mathrm{A}_t^{\mathrm{EDR}}(\mu_t,\alpha,\eta)$ and $\mathrm{H}_t^{\mathrm{EDR}}(\mu_t,\alpha,\eta)$ with

$$\widehat{\mathrm{A}_t^{\mathrm{EDR}}}(\mu_t,\alpha,\eta)=E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\frac{\partial\varphi_t\left(X,Y,T;\mu_t,\alpha,\eta\right)}{\partial\alpha'}\right]$$

$$\times E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\frac{\partial S_\alpha(T,X;\alpha)}{\partial\alpha'}\right]^{-1}$$

$$\widehat{\mathrm{H}_t^{\mathrm{EDR}}}(\mu_t,\alpha,\eta)=E_n\left[\frac{V^{(1)}}{q^{(1)}(S,Y,T)}\frac{\partial\varphi_t\left(X,Y,T;\mu_t,\alpha,\eta\right)}{\partial\eta'}\right]$$

$$\times E_n\left[\frac{V^{(1)}I_{\{T=t\}}}{q^{(1)}(S,Y,T)}\frac{\partial S_{t\eta}(Y,X;\eta)}{\partial\eta'}\right]^{-1},$$

respectively.

Since the covariates $W$ are valuable, it is important to incorporate the $W$'s collected in the intermediate step into the validation sampling at the second phase. Specifically, we propose the following two-step procedure:

- If $V^{(1)} = 1$, set $V = 1$; i.e. those subjects with $W$ measured in the intermediate step will directly be included as part of the validation sample at the second phase.

- If $V^{(1)} = 0$, generate, for subjects in stratum $k$ (i.e., $r(S, Y, T) = k$), $V$ with conditional probability $q_k^{(2)} = P[V = 1 | r(S, Y, T) = k, V^{(1)} = 0]$; i.e. the remaining subjects in the validation sample will be a random subset from those not selected at the intermediate step.

For subjects in stratum $k$,

$$q_k = q_k^{(1)} + q_k^{(2)} \cdot \left(1 - q_k^{(1)}\right). \tag{3.12}$$

Note that (3.12) implies that

$$q_k^{(1)} \leq q_k \leq 1, \tag{3.13}$$

where $q_k^{(1)}$ is pre-specified. Restriction (3.13) should be added to the constrained optimization to obtain the optimal $q_k$'s. The validation sampling probabilities in the second stage, $q_k^{(2)}$ can be determined from (3.12) as

$$q_k^{(2)} = \frac{q_k - q_k^{(1)}}{1 - q_k^{(1)}}. \tag{3.14}$$

**Summary**

In sum, we conduct optimization (3.9) with the following modifications:

- estimates of marginal probabilities are replaced by estimates computable from the first stage data
- conditional expectations and variances are replaced by estimates computable using intermediate stage data
- constraint (3.13) is added

The solution to the optimization can be used to obtain $q_k^{(2)}$ using (3.14). These stratum-specific probabilities can then be used to sample additional validation subjects. Inference can then proceed with the SIPW or EDR estimators with the sampling weights treated as known or estimated.

## 3.3 Simulation study

To illustrate the finite sample performance of our proposed procedure, we conducted a simulation study with 2000 runs, each with a sample size $n = 2000$.

We generated data under the following true data generating mechanism: $S = (S_1, S_2)'$, $S_1 \sim$ Bernoulli(0.5), $S_2 \sim$ Bernoulli(0.3) and $W \sim \mathcal{N}(0, 1)$. The three covariates were assumed to be independent. Given the complete covariate vector $X = (S_1, S_2, W)'$, $T$, $Y_1$ and $Y_0$ were generated as independent Bernoulli distributed random variables with probabilities

$$P[T = 1|X] = \text{expit}(S_1 + 0.2S_2 + 0.8W)$$

$$P[Y_1 = 1|X] = \text{expit}(S_1 + 0.3S_2 + W)$$

$$P[Y_0 = 1|X] = \text{expit}(1 + 1.5S_1 + 0.5S_2 + 1.5W).$$

We then set $Y = TY_1 + (1-T)Y_0$. Under this data generation scheme, $(S, Y, T)$ contains only categorical components, and the first phase data naturally form $K = 16$ strata.

We considered the following two scenarios for generating $V$ within each stratum so that approximately 50% of subjects from each treatment arm were selected in the second phase.

- "Equal number per stratum": $P[V = 1|S, Y, T] = q^{\text{eq}}(S, Y, T)$ so that the expected number of individuals selected per stratum is set equal to the minimum of the stratum size and the expected total selected divided by the number of strata $K$.

- "Optimal": $P[V = 1|S, Y, T] = q^{\text{opt}}(S, Y, T)$, with $V$ generated by the following 3 steps:

    1. Generate intermediate $V^{(1)}$ with probability $P[V^{(1)} = 1|S, Y, T] = q^{\text{eq}(1)}(S, Y, T)$ so that approximately 25% subjects from each treatment arm is selected in the intermediate step using the "equal number per stratum" rule above;

    2. Solve for the constrained optimal $q^{\text{opt}}(S, Y, T)$ minimizing the asymptotic variance of the estimation procedure of interest (SIPW or EDR; known or estimated weights);

    3. If $V^{(1)} = 1$, we set $V = 1$; else generate $V$ according to the conditional probability

$$P[V = 1|S, Y, T, V^{(1)} = 0] = \frac{q^{\text{opt}}(S, Y, T) - q^{\text{eq}(1)}(S, Y, T)}{1 - q^{\text{eq}(1)}(S, Y, T)}$$

We compared the performance of eight estimation procedures: two estimators (SIPW and EDR) each with four ways of handling $q(S, Y, T)$ ($q^{\text{eq}}$ known, $q^{\text{opt}}$ known, $q^{\text{eq}}$ estimated - $\widehat{q^{\text{eq}}}$ , $q^{\text{opt}}$ estimated- $\widehat{q^{\text{opt}}}$). We focused on the treatment effect $\Delta^* = \mu_1^* - \mu_0^*$. The corresponding estimator of $\Delta^*$ is $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$. When the weights are treated as known, the influence function for $\hat{\Delta}$ will take the form:

$$\text{IF}_1\left(O; \mu_1^*, \tilde{\alpha}, \tilde{\eta}\right) - \text{IF}_0\left(O; \mu_0^*, \tilde{\alpha}, \tilde{\eta}\right);$$

when the weights are estimated, the influence function for $\hat{\Delta}$ will take the form:

$$\text{IF}_1^{\dagger}\left(O; \mu_1^*, \tilde{\alpha}, \tilde{\eta}\right) - \text{IF}_0^{\dagger}\left(O; \mu_0^*, \tilde{\alpha}, \tilde{\eta}\right).$$

Methods developed in Section 3.2 are still applicable except that $h_{1t}(X, Y, T; \mu_t, \alpha, \eta)$ and $h_{2t}(S, Y, T; \mu_t, \alpha, \eta)$ are replaced by the corresponding differences

$$h_{11}(X, Y, T; \mu_1, \alpha, \eta) - h_{10}(X, Y, T; \mu_0, \alpha, \eta)$$

and

$$h_{21}(S, Y, T; \mu_1, \alpha, \eta) - h_{20}(S, Y, T; \mu_0, \alpha, \eta)$$

respectively. The corresponding conditional expectation and variances in the optimization are estimated based on the intermediate stage data using the same ideas discussed above.

For each intended estimation procedure, once the sampling scheme was determined in the second phase, we drew the validation sample accordingly, computed the corresponding estimator and estimated its standard error based on the influence function. We further constructed 95% Wald confidence intervals accordingly.

The true values $\mu_1^* = 0.6153$, $\mu_0^* = 0.7839$ and $\Delta^* = -0.1687$ were determined in one extra simulation with sample size $10^8$. In the same simulation, we also computed $E[Y|T = 1] = 0.6831$, $E[Y|T = 0] = 0.6782$, and $E[Y|T = 1] - E[Y|T = 0] = 0.0049$, which indicates that our data generating mechanism has strong selection bias.

Table 3.1: Simulation statistics for treatment effect $(\mu_1^* - \mu_0^*)$ estimators, under hypothetical two-phase sampling designs

| Estimation Procedure | Bias ($\times 10^{-3}$) | Mean SE ($\times 10^{-3}$) | SD ($\times 10^{-3}$) | Loss of efficiency$^{\dagger}$(%) | 95% CI coverage (%) | Convergence rate (%) |
|---|---|---|---|---|---|---|
| SIPW ($q^{\text{eq}}$) | -0.559 | 27.2 | 27.6 | 55.6 | 95.0 | 100.0 |
| SIPW ($q^{\text{opt}}$) | -1.346 | 23.7 | 24.2 | 42.2 | 94.3 | 100.0 |
| SIPW ($\widehat{q^{\text{eq}}}$) | -0.088 | 21.1 | 21.6 | 27.4 | 94.7 | 100.0 |
| SIPW ($\widehat{q^{\text{opt}}}$) | -1.808 | 19.5 | 20.0 | 15.4 | 94.3 | 94.8 |
| EDR ($q^{\text{eq}}$) | -0.417 | 20.8 | 21.3 | 25.4 | 94.3 | 100.0 |
| EDR ($q^{\text{opt}}$) | -0.016 | 19.1 | 19.8 | 13.6 | 94.3 | 99.7 |
| EDR ($\widehat{q^{\text{eq}}}$) | 0.005 | 20.8 | 21.3 | 25.4 | 94.5 | 100.0 |
| EDR ($\widehat{q^{\text{opt}}}$) | -2.233 | 19.1 | 19.3 | 9.1 | 93.9 | 99.4 |

$^{\dagger}$ Loss of efficiency (%): defined as (1-RE)$\times$100%, where the RE represents the relative efficiency with respect to the same estimator (SIPW or EDR) with $W$ collected on the entire sample, computed using ratio of the Monte Carlo variances.

The results of the simulation study are summarized in Table 3.1. Across the 2000 simulated datasets, we report the convergence rate (see seventh column); and among those converged, we report average bias, average of standard error estimate, Monte Carlo standard deviation of parameter estimates, loss of efficiency compared to the same estimator (SIPW or EDR) with $W$ collected on the entire sample, and coverage of 95% confidence intervals (second through sixth columns, respectively). With the exception of SIPW estimator with $\widehat{q^{\text{opt}}}$, the convergence rate was greater than 99%. The bias for all the eight procedures is small. Moreover, the average of standard error estimates (third column)

agrees well with the Monte Carlo standard deviation of the parameter estimators (fourth column), and the coverage of the estimated 95% confidence intervals is close to their nominal level.

In terms of efficiency, the EDR estimator is at least as efficient as the SIPW estimator. When $V$ is generated "optimally", we observe lower variability compared to when $V$ is generated using the "equal number per stratum" procedure. This is true for both SIPW and EDR estimators, regardless of whether the weights are known or estimated. This is particularly obvious when comparing SIPW ($q^{\text{opt}}$) to SIPW ($q^{\text{eq}}$) - 12% $\sim$ 13% reduction in variability. For SIPW, we can also see reduced variability when using estimated rather than known weights - there is a 22% reduction in variability for the "equal number per stratum" sampling procedure and a 17% for the "optimal" sampling procedure. For EDR, there is no reduction. This is because there is no model mis-specification, i.e. $\tilde{\alpha} = \alpha^*$, $\tilde{\eta} = \eta^*$. This implies that $\mathrm{A}_t^{\text{EDR}}(\mu_t^*, \alpha^*, \eta^*) = \mathrm{H}_t^{\text{EDR}}(\mu_t^*, \alpha^*, \eta^*) = 0$. Thus, the influence function when treating the weights as known or estimated are identical.

An important question is: how much efficiency is lost due to not collecting $W$ on the non-validation sample. To address this question, we compared the Monte Carlo variances between each estimation procedure under two phase sampling and the corresponding estimation procedure with $W$ collected on the entire sample. We computed the loss of efficiency (%), as shown in the fifth column of Table 3.1. For reference, the Monte Carlo standard deviations for the SIPW and EDR estimators were both estimated to be approximately 18.4 when $W$ was recorded on all subjects. For the SIPW estimator, the resulting loss of efficiency ranges from 15% to 55%. For the EDR estimator, the resulting loss of

efficiency is approximately 25% for the "equal number per stratum" sampling scheme and 10-15% for the "optimal" sampling scheme.

## 3.4 Data analysis

Right heart catheterization (RHC) is a diagnostic procedure to evaluate how well the heart is functioning in critically ill patients. The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) was an observational study in which data were collected on the outcomes, treatments and predictive factors of seriously ill, hospitalized adult patients at 5 medical center in the United States (Connors et al., 1996). We analyzed data on 5,735 SUPPORT patients, of which 2,184 (38.1%) had RHC within the first 24 hours after study entry. Overall, 66.6% patients survived beyond 30 days; 62.0% with RHC and 69.4% without RHC. This translates into a crude estimate of the difference in 30-day survival for RHC vs. no RHC of -7.36%, (95% CI: -9.90% to -4.83%; $P < .001$). Because of confounding by indication, it is important to estimate an adjusted effect of RHC on 30-day survival.

Connors et al. (1996), using an expert panel, identified a list of pre-treatment variables that were *a priori* considered to be predictive of the decision to treat a patient with RHC. These variables ($X$) included: age, sex, race, education, income, insurance, primary and secondary disease categories, admission diagnosis, ADL and DASI scores two weeks prior to admission, resuscitation order, cancer status, estimate of surviving 2 months, physiology component of APACHE III

score, Glasgow Coma score, weight, temperature, mean blood pressure, respiratory rate, heart rate, $PaO_2/FIO_2$ ratio, $PaCO_2$, $pH$, WBC count, hematocrit, sodium, potassium, creatinine, bilirubin, albumin, urine output, and co-morbidities. They used these variables to build a propensity score model (i.e., a model for $P[T = 1|X]$) and matched each patient with RHC, if possible, to a patient who did not receive RHC based on the estimated propensity score and disease category. Based on 1,008 matched pairs, they found that RHC have a lower 30-day survival than not performing RHC, with an odds ratio for mortality of 1.24 (95% CI: 1.03 to 1.49; $P = 0.03$).

Hirano and Imbens (2001) and Tan (2006) also analyzed these data to estimate the causal effect of RHC, i.e. difference of 30-day survival had all patients been treated with RHC vs. untreated with RHC. The two papers also found that RHC had significant decreased 30-day survival compared to not performing RHC. Hirano and Imbens (2001) used a combination of outcome regression adjustment and propensity score weighting. Under various model specifications, they reported causal effects ranging from -6.8% to -4.8% (with standard errors ranging from 1.2% to 1.6%). Under a propensity score model, Tan (2006) derived a local efficient estimator which is optimal under correct specification of an outcome regression model. He also developed a doubly robust version of this estimator which is consistent and asymptotically normal if either the the propensity score or outcome regression models are correctly specified. He carefully built a propensity score model for the RHC data, allowing for interactions of covariates. He applied the two proposed estimators and reported, under various model specifications, causal effect estimates ranging from -5.2% to -4.0% (with standard errors around 1.5% $\sim$ 1.6%).

To illustrate our proposed methods, we implemented a hypothetical two-phase sampling design on the RHC dataset. Like Hirano and Imbens (2001); Tan (2006), we are interested in the causal effect of RHC on 30-day survival.

In the first phase, apart from the 30-day survival indicator ($Y$) and the RHC treatment indicator ($T$), we considered the "inexpensive" covariates ($S$) to be collected on all subjects from the following demographic and socioeconomic variables: age, sex, race, education, income and insurance. To identify stratification factors, we fit a logistic regression with 30-day survival indicator as the outcome and the inexpensive covariate candidates plus the RHC treatment indicator as predictors. Age and income ( categorized as: under \$11k, \$11–\$25k, \$25–\$50k, over \$50k) were significant predictors at the 0.05 level.

Given the first phase data, we classified all subjects into 32 strata formed by dichotomized age (by median), four levels of income, 30-day survival indicator and RHC treatment indicator. As in the simulation study, we considered the "equal number per stratum" (called "equal" below) and the "optimal" validation sampling schemes. The goal was to sample approximately 50% of subjects within each treatment arm. We included 71 covariates that had no missing data on all 5,735 subjects. All covariates, except age and income level ($S$), were treated as expensive covariates ($W$).

We implemented the eight estimation procedures described in our simulation study. Logistic regression models were used to model $\pi_t^*(X)$, $\mu_1^*(X)$ and $\mu_0^*(X)$. For the sake of computational stability, all models included an intercept and the main effects of the 71 covariates (i.e., no interactions).

Our analysis results are summarized in Table 3.2. In addition to the eight estimation procedures under the hypothetical two-phase sampling design, we

Table 3.2: Causal effect of RHC on 30-day survival (%), under hypothetical two-phase sampling designs

| Estimation Procedure | ATE on survival (%) | SE | 95% CI | P-value |
|---|---|---|---|---|
| SIPW ($q^{\text{eq}}$) | -7.05 | 2.59 | (-12.12, -1.98) | 0.006 |
| SIPW ($q^{\text{opt}}$) | -6.88 | 2.07 | (-10.93, -2.83) | 0.001 |
| SIPW ($\widehat{q^{\text{eq}}}$) | -8.61 | 2.60 | (-13.70, -3.51) | 0.001 |
| SIPW ($\widehat{q^{\text{opt}}}$) | -5.23 | 1.95 | (-9.06, -1.40) | 0.007 |
| SIPW (1) | -5.50 | 1.53 | (-8.50, -2.50) | <0.001 |
| EDR ($q^{\text{eq}}$) | -7.43 | 2.02 | (-11.39, -3.47) | <0.001 |
| EDR ($q^{\text{opt}}$) | -6.38 | 1.83 | (-9.97, -2.78) | 0.001 |
| EDR ($\widehat{q^{\text{eq}}}$) | -7.95 | 2.19 | (-12.24, -3.66) | <0.001 |
| EDR ($\widehat{q^{\text{opt}}}$) | -5.46 | 1.83 | (-9.05, -1.88) | 0.003 |
| EDR (1) | -6.21 | 1.51 | (-9.16, -3.25) | <0.001 |

also show, for comparative purposes, the results of the SIPW and EDR estimators using the expensive covariates $W$ for the entire sample (i.e. sampling probability $q = 1$), these latter estimators are labeled as SIPW (1) and EDR (1). For each estimation procedure, we report the estimated causal effects of interest (%) and associated standard error (SE), the 95% Wald confidence interval and the two-sided p-value based on the Z-score.

For both the SIPW and EDR estimators, the estimated treatment effect is larger (in absolute value) and closer to the crude estimate (-7.36%) under the"equal" sampling scheme as compared to the "optimal" sampling scheme. The estimates under "optimal" sampling range from -6.9% to -5.2%, closer to the full estimates when $W$ is collected on the entire sample. Under "equal" sampling, the standard error is about 2.6% for the SIPW estimator and about 2.0% for EDR estimator. The comparable standard errors are lower under "optimal" sampling (2.0% and 1.8%, respectively) and full sampling (1.5% and

1.5%, respectively) . The SIPW and EDR confidence intervals under "optimal" sampling are 23% and 10% shorter than those under "equal" sampling, and 25% and 20% longer than those under full sampling. Estimation of the sampling weights did not meaningfully reduce the standard errors. The SIPW and EDR estimates and estimated standard errors under full sampling are comparable with the results in Hirano and Imbens (2001) and Tan (2006). All estimation procedures in Table 3.2 yield p-values less than 0.01, indicating significantly lower 30-day survival for RHC treatment.

## 3.5 Conclusion and Discussion

In this paper, we studied the outcome-dependent two-phase sampling design for estimating the causal effect of a treatment from observational data, and proposed an algorithm to find the optimal stratum-specific sampling probabilities for drawing the validation sample in the second phase. Our method aims at minimizing the variance for pre-specified type of estimators subject to the constraint of the validation sample size. To estimate the variance, we propose an intermediate step to collect information on the distribution of the expensive covariates conditional on the first phase data. An advantage of our methodology is that data collected in the intermediate step not only serves as a pilot data for estimation purposes, but also constitutes part of the final validation sample. Our procedure makes maximal use of the data and incorporates sampling uncertainty into the final estimation procedure. We also incorporated more efficient versions of the Wang et al. (2009) estimators that rely on estimates of the sampling probabilities. Our simulation study and empirical analysis demonstrated

that our "optimal" sampling strategy is more efficient than an "equal number per stratum" sampling strategy.

In this paper, we assumed that the stratification function was pre-specified. An open question is how to choose this function in order to further improve efficiency. This is particularly difficult when the first phase covariates are high-dimensional. We also pre-specified the intermediate validation sampling probabilities. While larger sampling probabilities will ensure a better estimate of the variance of a given estimator in the optimization, it constrains the amount of sampling at the second stage (since subjects selected at the intermediate step count toward the total number of validation subjects). It is an open question as to the proper balance between the number of subjects sampled at the intermediate and final steps.

The methods in this paper were developed within an independent and identically distributed (i.i.d.) observation framework. Some studies with two-phase sampling designs may employ binomial sampling within strata, i.e. selection of a specified number of subjects from each stratum. Future work could focus on inference and related optimal designs under binomial or other more complicated sampling schemes.

# Chapter 4

# Influence Function Based Fast Double Bootstrap Confidence Intervals

## 4.1 Introduction

Two-sided confidence intervals for a parameter of interest that are based on asymptotic normal approximations are second order accurate (i.e. with coverage error $O(n^{-1})$, where $n$ denotes sample size). Equal-tailed two-sided bootstrap confidence intervals, such as produced by bootstrap percentile, bootstrap-t (i.e. studentized bootstrap) and bootstrap $BC_a$ also have second order coverage error (Efron and Tibshirani, 1994; Tu and Shao, 1995; DiCiccio and Efron, 1996; Davison and Hinkley, 1997). These intervals can have poor coverage in small to medium sized samples.

The iterated or double bootstrap was proposed to reduce the coverage error (Hall, 1986). This procedure involves a nested level of resampling to calibrate the nominal level of the confidence interval. This idea has also been called bootstrap calibration (Efron and Tibshirani, 1994; Tu and Shao, 1995) and bootstrap prepivoting (Beran, 1987). The calibration operation can be repeatedly iterated

to obtain higher order accurate confidence intervals. Hall and Martin (1988) provided a unified account for the coverage correction of confidence intervals in a general framework of bootstrap iterations, and showed that each iteration reduces the coverage error by a factor of $O(n^{-1})$. In practice, employing more than one level of nested resampling is too computationally intensive to be of practical use. In this paper, we focus the double bootstrap.

Double bootstrap itself is computationally intensive and there have been a number of proposals to reduce its computational burden. Nankervis (2005) proposed stopping rules to reduce the number of inner level bootstrap replications. Davidson and MacKinnon (2007) and Giacomini et al. (2013) suggested the fast/warp-speed double bootstrap method that uses only a single simulation at the inner level of resampling. Recently, Chang and Hall (2015) pointed out that the coverage accuracy of the fast/warp-speed double bootstrap method is not improved over single bootstrap resampling. Approximations to inner level resampling have been investigated. Hall and Maesono (2000) introduced a weighted bootstrap resampling approach to reduce the computational burden while preserving the coverage accuracy. However, their method is complicated, and can still be computationally prohibitive in medium-sized samples. DiCiccio et al. (1992) introduced an easy and accurate saddlepoint approximation (Daniels, 1954, 1987; Diciccio and Martin, 1991) to the bootstrap distribution function that obviates the need for an inner level of resampling. A critical requirement underlying the utility of their approach is that the estimator for the parameter of interest is a smooth function of means of independent vectors.

In this paper, we extend the idea of DiCiccio et al. (1992) to more complicated settings where the parameter of interest is the solution to an estimating

function involve nuisance parameters. We propose to replace the original estimator or its pivot with the corresponding influence function in the resampling procedures, and calibrate the coverage error by using saddlepoint approximate double bootstrap approach. Our idea of "resampling influence functions" was inspired by Armstrong et al. (2014), where single layer bootstrap resampling of the influence function was proposed in similar settings, and the corresponding confidence intervals were showed to be consistent in coverage. Under mild regularity conditions, we further show the second order accuracy of equal-tail two-sided confidence intervals using their method, and improve the coverage accuracy to at least third order, i.e. with error of order $O(n^{-3/2})$, by saddlepoint approximate double bootstrap.

Our methods are tailored to the case where the estimator for the nuisance parameter is relatively hard to compute, but the influence function for the estimator of the parameter of interest can be easily obtained. We take full advantage of the properties of the influence function: firstly, estimators that admit an asymptotic linear representation allow saddlepoint approximation to the bootstrap resampling distribution; secondly, estimation of nuisance parameter is "borrowed" from the previous resampling level, and thus there is no need for it to be re-computed for each resample replication. Consequently, our fast double bootstrap algorithm is not only as speedy as most single layer bootstrap procedures, but with even better confidence interval coverage accuracy.

The paper is organized as follows. Section 4.2 introduces the framework and notation, gives a brief review of the iterated bootstrap procedures for constructing equal-tail two-sided confidence intervals, and proposes the algorithm for our influence function based fast double bootstrap interval. Section 4.3 establishes

the third order accuracy of the proposed fast double bootstrap interval. A simulation study in Section 4.4 demonstrates that, compared to other methods, our fast double bootstrap interval has desirable empirical coverage for small to medium sample sizes. The final section is devoted to a discussion.

## 4.2 Framework and fast double bootstrap

Assume we have data from a random sample $\mathcal{X} = (X_1, \ldots, X_n)$ drawn from some unknown distribution $F$ indexed by a (scalar) parameter of interest $\mu$ and a multi-dimensional nuisance parameter $\gamma$ which belongs to a Banach space $\mathcal{H}$ with a norm $\|\cdot\|$. To estimate $\mu$, consider some smooth, unbiased estimating function $\psi(x; \mu, \gamma)$ such that

$$E\left[\psi(X; \mu^\dagger, \gamma^\dagger)\right] = 0,$$

where $\mu^\dagger$ and $\gamma^\dagger$ are the true values for $\mu$ and $\gamma$, respectively. Suppose the estimator $\hat{\mu}$ is defined as the solution to the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i; \mu, \hat{\gamma}(\mu)) = 0, \tag{4.1}$$

where $\hat{\gamma}(\mu)$ is a profile estimator of $\gamma$ based on sample $\mathcal{X}$, for given $\mu$. Assume $\hat{\gamma}(\mu)$ is a smooth function of $\mu$, and $\hat{\gamma}(\mu^\dagger)$ is $\sqrt{n}-$consistent. We are interested in constructing an equal-tail two-sided bootstrap confidence interval (CIs) for $\mu$.

Let $\mathcal{X}^*$ denote a generic resample from $\mathcal{X}$, and $\mathcal{X}^{**}$ denote a resample from $\mathcal{X}^*$. Suppose we have $B$ resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$ based on the $\mathcal{X}$, and $C$ second level resamples $\mathcal{X}_{b1}^{**}, \ldots, \mathcal{X}_{bC}^{**}$ based on each bootstrap resample $\mathcal{X}_b^*$ ($b = 1, \ldots, B$). Furthermore, for given $\mu$, let $\hat{\gamma}^*(\mu)$ and $\hat{\gamma}^{**}(\mu)$ represent the versions

of $\hat{\gamma}(\mu)$ computed using $\mathcal{X}^*$ and $\mathcal{X}^{**}$, respectively, instead of $\mathcal{X}$. Specifically, $\hat{\mu}^*$ and $\hat{\mu}^{**}$are the solution to $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i^*;\mu,\hat{\gamma}^*(\mu))=0$ and $\frac{1}{n}\sum_{i=1}^{n}\psi(X_i^{**};\mu,\hat{\gamma}^{**}(\mu))=0$, respectively. We write $\hat{\mu}_b^*$ and $\hat{\mu}_{bc}^{**}$ to denote the corresponding versions computed using $\mathcal{X}_b^*$ and $\mathcal{X}_{bc}^{**}$ $(b=1,\ldots,B;\ c=1,\ldots,C)$, respectively. In addition, $\gamma$ can be estimated by $\hat{\gamma}(\hat{\mu})$, $\hat{\gamma}^*(\hat{\mu}^*)$ and $\hat{\gamma}^{**}(\hat{\mu}^{**})$ based on $\mathcal{X}$, $\mathcal{X}^*$ and $\mathcal{X}^{**}$, respectively. For simplicity, unless noted otherwise, we shall suppress such dependence on the estimator of $\mu$ and write $\hat{\gamma}$, $\hat{\gamma}^*$ and $\hat{\gamma}^{**}$ to indicate $\hat{\gamma}(\hat{\mu})$, $\hat{\gamma}^*(\hat{\mu}^*)$ and $\hat{\gamma}^{**}(\hat{\mu}^{**})$ respectively.

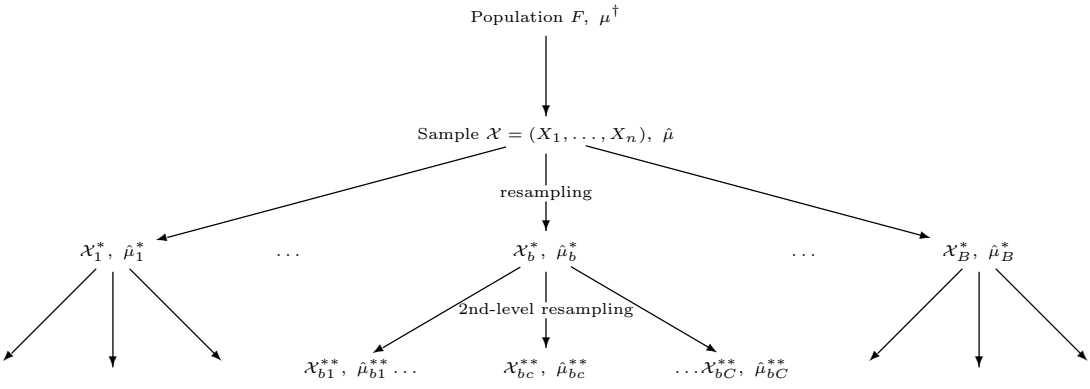



Figure 4.1: Diagram of double (iterated) bootstrap

The diagram in Figure 4.1 outlines the structure of double (iterated) bootstrap resampling. Following DiCiccio et al. (1992), we hereby summarize the double (iterated) bootstrap CI algorithm with coverage correction as follows:

- Denote the uncorrected CI of nominal coverage $1-\alpha$ by $I_0(\alpha;\mathcal{X},\mathcal{X}^*)$, e.g. the percentile method interval $\left[\tilde{Q}^*(\alpha/2),\tilde{Q}^*(1-\alpha/2)\right]$, where $\tilde{Q}^*(x)$ is the $x$-level quantile of the distribution of $\hat{\mu}^*$ given $\mathcal{X}$, $0<\alpha<1$. Practically, we plug in the empirical quantiles of $\{\hat{\mu}_b^*;\ b=1,\ldots,B.\}$

- Define the coverage probability $\pi(\alpha) := P\left[\mu^{\dagger} \in I_0\left(\alpha; \mathcal{X}, \mathcal{X}^*\right)\right]$.

- The bootstrap estimate of $\pi(\alpha)$ is computed by taking the sample $\mathcal{X}$ as the population, using $\mathcal{X}^*$ and $\mathcal{X}^{**}$ in place of $\mathcal{X}$ and $\mathcal{X}^*$, respectively. Practically, we plug in empirical average for probability where necessary to obtain:

$$\begin{aligned}
\hat{\pi}(\alpha) &= P\left[\hat{\mu} \in I_0\left(\alpha; \mathcal{X}^*, \mathcal{X}^{**}\right) \middle| \mathcal{X}\right] \\
&\approx \frac{1}{B}\sum_{b=1}^{B} I\left\{\tilde{Q}_b^{**}\left(\frac{\alpha}{2}\right) \leq \hat{\mu} \leq \tilde{Q}_b^{**}\left(1-\frac{\alpha}{2}\right)\right\} \qquad (4.2) \\
&= \frac{1}{B}\sum_{b=1}^{B} I\left\{\frac{\alpha}{2} \leq P\left[\hat{\mu}^{**} \leq \hat{\mu} \middle| \mathcal{X}_b^*\right] \leq 1-\frac{\alpha}{2}\right\} \qquad (4.3) \\
&\approx \frac{1}{B}\sum_{b=1}^{B} I\left\{\frac{\alpha}{2} \leq \frac{1}{C}\sum_{c=1}^{C} I\left\{\hat{\mu}_{bc}^{**} \leq \hat{\mu}\right\} \leq 1-\frac{\alpha}{2}\right\}, \qquad (4.4)
\end{aligned}$$

  where $\tilde{Q}_b^{**}(x)$ in (4.2) is the $x$-level quantile of the distribution of $\hat{\mu}^{**}$ given $\mathcal{X}_b^*$.

- Note that the interval $I_0\left(\alpha + \delta_n; \mathcal{X}, \mathcal{X}^*\right)$, where $\pi(\alpha + \delta_n) = 1 - \alpha$, has the exact coverage $1 - \alpha$. We then find the bootstrap estimate $\hat{\delta}_n$ for $\delta_n$, as the solution to

$$\hat{\pi}(\alpha + \hat{\delta}_n) = 1 - \alpha.$$

- The double (iterated) bootstrap CI for $\mu$ is

$$I_1(\alpha; \mathcal{X}, \mathcal{X}^*, \mathcal{X}^{**}) = I_0(\alpha + \hat{\delta}_n; \mathcal{X}, \mathcal{X}^*).$$

To reduce the computational demands, DiCiccio et al. (1992) suggested that the saddlepoint approximation to the cumulative distribution function (CDF)

$P\left[\hat{\mu}^{**} \leq \cdot \,|\, \mathcal{X}_b^*\right]$ in (4.3) should be employed to replace the inner level of resampling in (4.4). This "shortcut" only works for estimators which can be represented as a smooth function of means of independent vectors. Note, however, that our estimator $\hat{\mu}$ is defined indirectly by solving the estimating equation (4.1) which itself is the mean of dependent functions $\{\psi(X_i; \mu, \hat{\gamma}(\mu)); i = 1, \ldots, n\}$ sharing the profile estimator $\hat{\gamma}(\mu)$. This motivates us to incorporate the asymptotic linear representation of the estimator, i.e. the influence function, into the resampling procedures.

Assume the influence function $\mathrm{IF}(X; \mu^\dagger, \gamma^\dagger)$ for $\hat{\mu}$ is readily available in our context. Under mild regularity conditions, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\mu} - \mu^\dagger\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(X_i; \mu^\dagger, \gamma^\dagger) + o_P(1) \\ &= \sqrt{n} \cdot \overline{\mathrm{IF}}\left(\mathcal{X}\right) + o_P(1), \end{aligned} \tag{4.5}$$

where

$$\overline{\mathrm{IF}}\left(\mathcal{X}\right) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}(X_i; \mu^\dagger, \gamma^\dagger)$$

Notice that $\overline{\mathrm{IF}}\left(\mathcal{X}\right)$ is the mean of independent and identically distributed random variables, and thus can be used as an ideal substitute for the "centered estimator" $\hat{\mu} - \mu^\dagger$ in the resampling procedures. Further, note that the influence function may involve expectations with respect to the population distribution.

Define the following "bootstrap versions" of $\overline{\mathrm{IF}}\left(\mathcal{X}\right)$:

$$\widehat{\overline{\mathrm{IF}}}(\mathcal{X}^*) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathrm{IF}}(X_i^*; \hat{\mu}, \hat{\gamma}) \tag{4.6}$$

$$\widehat{\overline{\mathrm{IF}}}^*(\mathcal{X}^{**}) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathrm{IF}}^*(X_i^{**}; \hat{\mu}^*, \hat{\gamma}^*), \tag{4.7}$$

where $\widehat{\mathrm{IF}}(\cdot)$ and $\widehat{\mathrm{IF}}^*(\cdot)$ are the versions of $\mathrm{IF}(\cdot)$ in which expectations are computed conditional on $\mathcal{X}$ and $\mathcal{X}^*$, respectively. Intuitively, if we start with the pivot $\sqrt{n}(\hat{\mu} - \mu^\dagger)$, its bootstrap counterpart $\sqrt{n}(\hat{\mu}^* - \hat{\mu})$ may be replaced by $\sqrt{n} \cdot \widehat{\overline{\mathrm{IF}}}(\mathcal{X}^*)$. The corresponding bootstrap confidence interval coverage can be estimated and calibrated by double bootstrap concepts where the inner level resampling is not actually necessary: the distribution of $\widehat{\overline{\mathrm{IF}}}^*(\mathcal{X}^{**})$ can be estimated by a saddlepoint approximation to its density given $\mathcal{X}^*$. Note that in (4.7), $\hat{\mu}^*$ and $\hat{\gamma}^*$ are fixed conditional on $\mathcal{X}^*$, and

$$\widehat{\mathrm{IF}}^*(X_1^{**}; \hat{\mu}^*, \hat{\gamma}^*), \ldots, \widehat{\mathrm{IF}}^*(X_n^{**}; \hat{\mu}^*, \hat{\gamma}^*)$$

are independent and identically distributed from the empirical distribution based on $\left\{\widehat{\mathrm{IF}}^*(X_i^*; \hat{\mu}^*, \hat{\gamma}^*);\ i = 1, \ldots, n\right\}$. We employ the following approximation formula suggested by Daniels (1987) and Diciccio and Martin (1991) to the tail probability

$$P\left[\widehat{\overline{\mathrm{IF}}}^*(\mathcal{X}^{**}) \leq x \middle| \mathcal{X}^*\right] = \Phi\left(r_x^*\right) + \phi\left(r_x^*\right)\left[\frac{1}{r_x^*} - \frac{1}{w_x^*}\right] + O_{P^*}\left(n^{-3/2}\right) \tag{4.8}$$

for $x$ over the range $x - n^{-1}\sum_{i=1}^{n} \widehat{\mathrm{IF}}^*(X_i^*; \hat{\mu}^*, \hat{\gamma}^*) = O_{P^*}\left(n^{-1/2}\right)$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution and density functions respectively,

$$r_x^* = \mathrm{sgn}\left(T_x^*\right)\sqrt{2n\left\{T_x^* x - K^*\left(T_x^*\right)\right\}}$$

$$w_x^* = \hat{T}_x\sqrt{nK^{*\prime\prime}\left(T_x^*\right)}$$

$$\begin{aligned} K^*(T) &= \log E\left[\exp\left\{T \cdot \widehat{\mathrm{IF}}^*(X^{**}; \hat{\mu}^*, \hat{\gamma}^*)\right\} \middle| \mathcal{X}^*\right] \\ &= \log\left\{\frac{1}{n}\sum_{i=1}^{n} \exp\left\{T \cdot \widehat{\mathrm{IF}}^*(X_i^*; \hat{\mu}^*, \hat{\gamma}^*)\right\}\right\} \end{aligned}$$

($K^*(T)$ is the cumulant generating function of $\widehat{\text{IF}}^*(X^{**}; \hat{\mu}^*, \hat{\gamma}^*)$ given $\mathcal{X}^*$), and the saddlepoint $T_x^*$ is defined by $K^{*\prime}(T_x^*) = x$. Note that in (4.8) above and subsequently, we refer to $P^*$ as the bootstrap distribution conditional on $\mathcal{X}$.

Using this approximation, we revise the previous double bootstrap CI algorithm accordingly, and propose the following influence function based fast double bootstrap CI algorithm:

- Define the influence function based single bootstrap CI of nominal coverage $1-\alpha$ as $I_0(\alpha; \mathcal{X}, \mathcal{X}^*) = [\hat{\mu} - Q^*(1-\alpha/2), \hat{\mu} - Q^*(\alpha/2)]$, where $Q^*(x)$ is the $x$-level quantile of $\widehat{\overline{\text{IF}}}(\mathcal{X}^*)$ given $\mathcal{X}$, $0 < \alpha < 1$. Practically, we plug in the empirical quantiles of $\left\{\widehat{\overline{\text{IF}}}(\mathcal{X}_b^*);\ b = 1, \ldots, B.\right\}$

- Define the coverage probability $\pi(\alpha) := P\left[\mu^\dagger \in I_0(\alpha; \mathcal{X}, \mathcal{X}^*)\right]$.

- The bootstrap estimate of $\pi(\alpha)$ is computed by taking the sample $\mathcal{X}$ as the population, using $\mathcal{X}^*$ and $\mathcal{X}^{**}$ in place of $\mathcal{X}$ and $\mathcal{X}^*$, respectively. Practically, we plug in empirical average for probability where necessary to obtain:

$$\begin{aligned}
\hat{\pi}(\alpha) &= P\left[\hat{\mu} \in I_0(\alpha; \mathcal{X}^*, \mathcal{X}^{**}) \middle| \mathcal{X}\right] \\
&\approx \frac{1}{B}\sum_{b=1}^{B} I\left\{\hat{\mu}_b^* - Q_b^{**}\left(1-\frac{\alpha}{2}\right) \le \hat{\mu} \le \hat{\mu}_b^* - Q_b^{**}\left(\frac{\alpha}{2}\right)\right\} \qquad (4.9) \\
&= \frac{1}{B}\sum_{b=1}^{B} I\left\{\frac{\alpha}{2} \le P\left[\widehat{\overline{\text{IF}}}_b^*(\mathcal{X}^{**}) \le \hat{\mu}_b^* - \hat{\mu}\middle| \mathcal{X}_b^*\right] \le 1-\frac{\alpha}{2}\right\} \qquad (4.10)
\end{aligned}$$

where $Q_b^{**}(x)$ in (4.9) is the $x$-level quantile of the conditional distribution of $\widehat{\overline{\text{IF}}}_b^*(\mathcal{X}^{**})$ given $\mathcal{X}_b^*$ and $\widehat{\overline{\text{IF}}}_b^*(\cdot)$ is the version of $\widehat{\overline{\text{IF}}}^*(\cdot)$ based on $\mathcal{X}_b^*$. The conditional probability of $\widehat{\overline{\text{IF}}}_b^*(\mathcal{X}^{**})$ in (4.10) can be estimated using the

saddlepoint approximation formula (4.8). Let $\tilde{\pi}(\alpha)$ denote the resulting approximation for $\hat{\pi}(\alpha)$.

- As in the previous version of the algorithm, interval $I_0\left(\alpha+\delta_n; \mathcal{X}, \mathcal{X}^*\right)$, where $\pi(\alpha+\delta_n) = 1-\alpha$, has the exact coverage $1-\alpha$. We then find the bootstrap estimate $\tilde{\delta}_n$ for $\delta_n$, as the solution to

$$\tilde{\pi}(\alpha+\tilde{\delta}_n) = 1-\alpha. \tag{4.11}$$

- The fast double bootstrap CI for $\mu$ is $I_0(\alpha+\tilde{\delta}_n; \mathcal{X}, \mathcal{X}^*)$.

## 4.3 Asymptotic coverage accuracy

We evaluate the coverage error of confidence intervals by the asymptotic accuracy, i.e. the convergence rate of the coverage probabilities of the confidence intervals to the nominal level. A confidence interval $I(\alpha; \mathcal{X})$ of $\mu^\dagger$ with nominal coverage $1-\alpha$ is said to be $k$th-order (asymptotically) accurate if

$$P\left[\mu^\dagger \in I\left(\alpha; \mathcal{X}\right)\right] - (1-\alpha) = O\left(n^{-k/2}\right).$$

In this section, we derive the third order accuracy of our proposed fast double bootstrap confidence interval.

As a prerequisite, we first establish the second order accuracy of the single layer, uncorrected bootstrap confidence interval
$I_0\left(\alpha; \mathcal{X}, \mathcal{X}^*\right) = [\hat{\mu} - Q^*(1-\alpha/2), \hat{\mu} - Q^*(\alpha/2)]$, where $Q^*(x)$ is the $x$-level quantile of the conditional distribution of $\widehat{\mathrm{IF}}(\mathcal{X}^*)$ given $\mathcal{X}$. Here we follow the assumptions and logic in Hall (1992) to derive the key points of the proof.

Given (4.5), the asymptotic variance of the pivot $\sqrt{n}(\hat{\mu}-\mu^\dagger)$ is

$$\sigma^{\dagger 2} = E\left[\mathrm{IF}(X_i; \mu^\dagger, \gamma^\dagger)^2\right]$$

which can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{IF}}(X_i; \hat{\mu}, \hat{\gamma})^2,$$

As above, let $\hat{\sigma}^{*2}$ and $\hat{\sigma}^{**2}$ denote the bootstrap version of $\hat{\sigma}^2$ computed using $\mathcal{X}^*$ and $\mathcal{X}^{**}$, respectively, instead of $\mathcal{X}$. We assume the distribution of the standardized pivot $\sqrt{n} \cdot \overline{\mathrm{IF}}(\mathcal{X})/\sigma^\dagger$ and the distribution of the studentized pivot $\sqrt{n}(\hat{\mu} - \mu^\dagger)/\hat{\sigma}$ admit the following Edgeworth expansions

$$P\left[\frac{\sqrt{n} \cdot \overline{\mathrm{IF}}\left(\mathcal{X}\right)}{\sigma^\dagger} \leq x\right] = \Phi(x) + \sum_{j=1}^{2} \frac{u_j(x)\phi(x)}{n^{j/2}} + O\left(n^{-3/2}\right) \tag{4.12}$$

$$P\left[\frac{\sqrt{n}(\hat{\mu} - \mu^\dagger)}{\hat{\sigma}} \leq x\right] = \Phi(x) + \sum_{j=1}^{2} \frac{\tilde{u}_j(x)\phi(x)}{n^{j/2}} + O\left(n^{-3/2}\right), \tag{4.13}$$

uniformly in $x$, where $u_j(x)$ and $\tilde{u}_j(x)$ are odd (even) polynomials for even (odd) $j$. Moreover, suppose the following (inverse) Cornish-Fisher expansions hold:

$$Q_s\left(y\right) = z_y + \sum_{j=1}^{2} \frac{v_j(z_y)}{n^{j/2}} + O\left(n^{-3/2}\right) \tag{4.14}$$

$$\tilde{Q}_s\left(y\right) = z_y + \sum_{j=1}^{2} \frac{\tilde{v}_j(z_y)}{n^{j/2}} + O\left(n^{-3/2}\right) \tag{4.15}$$

uniformly in $\epsilon < y < 1 - \epsilon$ for any $0 < \epsilon < \frac{1}{2}$, where $z_y = \Phi^{-1}(y)$, $Q_s(y)$ and $\tilde{Q}_s(y)$ are the $y$-level quantiles of the distribution of the pivots $\sqrt{n} \cdot \overline{\mathrm{IF}}(\mathcal{X})/\sigma^\dagger$ and $\sqrt{n}(\hat{\mu} - \mu^\dagger)/\hat{\sigma}$ respectively, and $v_j(x)$ and $\tilde{v}_j(x)$ are odd (even) polynomials for even (odd) $j$. In particular, we have

$$v_1(x) = -u_1(x) \quad \text{and} \quad v_2(x) = u_1(x)u_1'(x) - \frac{1}{2}xu_1^2(x) - u_2(x); \tag{4.16}$$

the formulae in (4.16) also hold when $u_j, v_j$ are replaced by $\tilde{u}_j, \tilde{v}_j$ respectively.

The coefficients in the polynomials $u_j, \tilde{u}_j, v_j, \tilde{v}_j$ are functions of expectations with respect to the population distribution. Write $\widehat{u_j}, \widehat{\tilde{u}_j}, \widehat{v_j}, \widehat{\tilde{v}_j}$ for the corresponding versions in which these expectations are computed conditional on the sample $\mathcal{X}$. Under regularity conditions, we have the bootstrap version of the quantile expansion (4.14):

$$Q_s^*(y) = z_y + \sum_{j=1}^{2} \frac{\widehat{v_j}(z_y)}{n^{j/2}} + O_P\left(n^{-3/2}\right), \tag{4.17}$$

where $Q_s^*(y)$ is the $y$-level quantile of the conditional distribution $\sqrt{n} \cdot \widehat{\overline{\mathrm{IF}}}(\mathcal{X}^*)/\hat{\sigma}$ given $\mathcal{X}$. Meanwhile, we have for each $0 < y < 1$,

$$\widehat{v_j}(z_y) - v_j(z_y) = O_P\left(n^{-1/2}\right), \quad j = 1, 2.$$

Consequently, we can show that for any $0 < \beta < 1$:

$$\begin{aligned}
P\left[\hat{\mu} - Q^*(\beta) \leq \mu^\dagger\right] &= P\left[\frac{\sqrt{n}(\hat{\mu} - \mu^\dagger)}{\hat{\sigma}} \leq \frac{\sqrt{n}Q^*(\beta)}{\hat{\sigma}}\right] \\
&= P\left[\frac{\sqrt{n}(\hat{\mu} - \mu^\dagger)}{\hat{\sigma}} \leq Q_s^*(\beta)\right] \\
&= P\left[\frac{\sqrt{n}(\hat{\mu} - \mu^\dagger)}{\hat{\sigma}} \leq z_\beta + \sum_{j=1}^{2} \frac{\widehat{v_j}(z_\beta)}{n^{j/2}}\right] + O\left(n^{-3/2}\right) \\
&= P\left[\frac{\sqrt{n}(\hat{\mu} - \mu^\dagger)}{\hat{\sigma}} \leq z_\beta + \sum_{j=1}^{2} \frac{v_j(z_\beta)}{n^{j/2}}\right] + \frac{p(z_\beta)\phi(z_\beta)}{n} + O\left(n^{-3/2}\right) \\
&= \beta + \frac{1}{\sqrt{n}} r_1(z_\beta)\phi(z_\beta) + \frac{1}{n} r_2(z_\beta)\phi(z_\beta) + O\left(n^{-3/2}\right),
\end{aligned} \tag{4.18}$$

where $p(x)$ is a polynomial whose coefficients are functions of expectations with

respect to the population distribution, and

$$r_1(x) = \tilde{u}_1(x) + v_1(x), \tag{4.19}$$

$$r_2(x) = \tilde{u}_2(x) + v_2(x) - \frac{1}{2}xv_1^2(x) + v_1(x)\left[\tilde{u}_1'(x) - x\tilde{u}_1(x)\right] + p(x). \tag{4.20}$$

The existence of $p(x)$ requires some algebra (omitted) and the second to last equality in (4.18) can be verified by a complicated argument presented in Hall (1992). The last equality (4.18) follows from applying the Edgeworth expansion (4.13) at $x = z_\beta + \sum_{j=1}^{2} n^{-j/2} v_j(z_\beta)$ and employing a Taylor series expansion along with the delta method to obtain the remainder of order $n^{-3/2}$. It is useful to notice that $r_1(x)$ is an even polynomial, since $\tilde{u}_1$ and $v_1$ are both even.

These results can be used to show that the influence function based single layer bootstrap confidence interval $I_0\left(\alpha; \mathcal{X}, \mathcal{X}^*\right) = [\hat{\mu} - Q^*(1-\alpha/2), \hat{\mu} - Q^*(\alpha/2)]$ has second order accurate coverage. Specifically,

$$\begin{aligned}
P\left[\mu^\dagger \in I_0\left(\alpha; \mathcal{X}, \mathcal{X}^*\right)\right] &= P\left[\hat{\mu} - Q^*\left(1 - \frac{\alpha}{2}\right) \leq \mu^\dagger \leq \hat{\mu} - Q^*\left(\frac{\alpha}{2}\right)\right] \\
&= P\left[\hat{\mu} - Q^*\left(1 - \frac{\alpha}{2}\right) \leq \mu^\dagger\right] - P\left[\hat{\mu} - Q^*\left(\frac{\alpha}{2}\right) < \mu^\dagger\right] \\
&= 1 - \frac{\alpha}{2} + \frac{1}{\sqrt{n}} r_1(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) + \frac{1}{n} r_2(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) \\
&\quad - \frac{\alpha}{2} - \frac{1}{\sqrt{n}} r_1(z_{\alpha/2})\phi(z_{\alpha/2}) - \frac{1}{n} r_2(z_{\alpha/2})\phi(z_{\alpha/2}) + O\left(n^{-3/2}\right) \\
&= 1 - \alpha + \frac{1}{n}\left[r_2(-z_{\alpha/2}) - r_2(z_{\alpha/2})\right]\phi(z_{\alpha/2}) + O\left(n^{-3/2}\right) \qquad (4.21) \\
&= 1 - \alpha + O(n^{-1})
\end{aligned}$$

where the third equality follows by (4.18), the fourth equality follows because

$r_1(\cdot)$ is an even function and the fifth equality follows because

$$c_0 = \left[r_2(-z_{\alpha/2}) - r_2(z_{\alpha/2})\right] \phi(z_{\alpha/2}) \tag{4.22}$$

is a constant.

Next, we discuss the impact of bootstrap iteration for calibration and the saddlepoint approximation on the coverage accuracy. To address this issue, we use the general iterative bootstrap framework introduced by Hall and Martin (1988). To start, define

$$f_{t_1,t_2}(\hat{\mu}; \mu^\dagger) = I\left\{\hat{\mu} - t_1 \leq \mu^\dagger \leq \hat{\mu} - t_2\right\} - (1-\alpha)$$

as a functional indexed by the pair $(t_1, t_2)$. Constructing a $(1-\alpha)$-level equal-tail two-sided confidence interval for $\mu^\dagger$ can be re-formulated as a problem of finding

$$\vec{t}(\beta) = \left(t\left(1 - \frac{\beta}{2}\right), t\left(\frac{\beta}{2}\right)\right) \tag{4.23}$$

so that

$$E\left[f_{\vec{t}(\beta)}(\hat{\mu}; \mu^\dagger)\right] = 0,$$

where $t(x)$ is the $x$-level quantile of some distribution; $0 < \alpha, \beta < 1$.

Consider the influence function based single layer bootstrap confidence interval:

$$\overrightarrow{Q^*}(\alpha) = \left(Q^*\left(1 - \frac{\alpha}{2}\right), Q^*\left(\frac{\alpha}{2}\right)\right),$$

where $Q^*(x)$ is the $x$-level quantile of $\widehat{\overline{\mathrm{IF}}}(\mathcal{X}^*)$ given $\mathcal{X}$. By (4.21), the associated coverage error can be re-expressed as

$$\pi(\alpha) - (1-\alpha) = E\left[f_{\overrightarrow{Q^*}(\alpha)}(\hat{\mu}; \mu^\dagger)\right] = c_0 n^{-1} + O\left(n^{-3/2}\right) \tag{4.24}$$

The idea of coverage correction for this interval is to calibrate the level $\alpha$ in $\overrightarrow{Q^*}(\alpha)$ so that the error is reduced to zero:

$$\pi(\alpha+\delta)-(1-\alpha)=E\left[f_{\overrightarrow{Q^*}(\alpha+\delta)}(\hat{\mu};\mu^\dagger)\right]=0 \tag{4.25}$$

In practice, as stated in (4.11), we solve the bootstrap version of (4.25), i.e.,

$$\hat{\pi}(\alpha+\delta)-(1-\alpha)=E\left[f_{\overrightarrow{Q^{**}}(\alpha+\delta)}(\hat{\mu}^*;\hat{\mu})\middle|\,\mathcal{X}\right]=0, \tag{4.26}$$

where

$$\overrightarrow{Q^{**}}(\beta)=\left(Q^{**}\left(1-\frac{\beta}{2}\right),Q^{**}\left(\frac{\beta}{2}\right)\right),$$

and $Q^{**}(x)$ is the $x$-level quantile of the conditional distribution of $\widehat{\overline{\mathrm{IF}}}^*(\mathcal{X}^{**})$ given $\mathcal{X}^*$. Following Hall and Martin (1988), define

$$d_n:=\frac{\partial}{\partial\delta}E\left[f_{\overrightarrow{Q^*}(\alpha+\delta)}(\hat{\mu};\mu^\dagger)\right]\Big|_{\delta=0},$$

and assume $d_n$ converges to a nonzero constant as $n\to\infty$. Let $\hat{c}_0$ and $\hat{d}_n$ denote the bootstrap estimates for $c_0$ (see (4.22)) and $d_n$ respectively, conditional on the sample $\mathcal{X}$. Then

$$\hat{\Delta}_n:=\sqrt{n}\left(\hat{c}_0\hat{d}_n^{-1}-c_0d_n^{-1}\right)=O_P(1).$$

Given (4.24) and performing a Taylor expansion of (4.25) around $\delta=0$,

$$E\left[f_{\overrightarrow{Q^*}(\alpha+\delta)}(\hat{\mu};\mu^\dagger)\right]=c_0n^{-1}+d_n\delta+O\left(n^{-3/2}\right).$$

Therefore, the solution to (4.25) is

$$\delta_n=-c_0d_n^{-1}n^{-1}+O\left(n^{-3/2}\right),$$

and similarly, the solution to (4.26) is

$$\hat{\delta}_n=-\hat{c}_0\hat{d}_n^{-1}n^{-1}+O_P\left(n^{-3/2}\right). \tag{4.27}$$

Rather than use $\hat{\pi}$ to do the calibration, we use $\tilde{\pi}$ which is found by plugging in the saddlepoint approximation to the conditional probability in (4.10). We have the following relationship between $\tilde{\pi}$ and $\hat{\pi}$:

$$
\begin{aligned}
&\tilde{\pi}(\alpha+\delta)-(1-\alpha) \\
&:= P^*\left[\frac{\alpha+\delta}{2} \leq P^{**}\left[\widehat{\overline{\mathrm{IF}}}^*\left(\mathcal{X}^{**}\right) \leq \hat{\mu}^*-\hat{\mu}\right]+O_{P^*}\left(n^{-3/2}\right) \leq 1-\frac{\alpha+\delta}{2}\right]-(1-\alpha) \\
&= P^*\left[\frac{\alpha+\delta}{2} \leq P^{**}\left[\frac{\sqrt{n} \cdot \widehat{\overline{\mathrm{IF}}}^*\left(\mathcal{X}^{**}\right)}{\hat{\sigma}^*} \leq \frac{\sqrt{n}\left(\hat{\mu}^*-\hat{\mu}\right)}{\hat{\sigma}^*}\right]+O_{P^*}\left(n^{-3/2}\right) \leq 1-\frac{\alpha+\delta}{2}\right] \\
&\quad-(1-\alpha) \\
&= P^*\left[\frac{\sqrt{n}\left(\hat{\mu}^*-\hat{\mu}\right)}{\hat{\sigma}^*} \leq Q_s^{**}\left(1-\frac{\alpha+\delta}{2}\right)+O_{P^*}\left(n^{-3/2}\right)\right] \\
&\quad-P^*\left[\frac{\sqrt{n}\left(\hat{\mu}^*-\hat{\mu}\right)}{\hat{\sigma}^*}<Q_s^{**}\left(\frac{\alpha+\delta}{2}\right)+O_{P^*}\left(n^{-3/2}\right)\right]-(1-\alpha) \\
&= P^*\left[Q_s^{**}\left(\frac{\alpha+\delta}{2}\right) \leq \frac{\sqrt{n}\left(\hat{\mu}^*-\hat{\mu}\right)}{\hat{\sigma}^*} \leq Q_s^{**}\left(1-\frac{\alpha+\delta}{2}\right)\right]-(1-\alpha)+O_P\left(n^{-3/2}\right) \\
&= P^*\left[\hat{\mu}^*-Q^{**}\left(1-\frac{\alpha+\delta}{2}\right) \leq \hat{\mu} \leq \hat{\mu}^*-Q^{**}\left(\frac{\alpha+\delta}{2}\right)\right]-(1-\alpha)+O_P\left(n^{-3/2}\right) \\
&= \hat{\pi}(\alpha+\delta)-(1-\alpha)+O_P\left(n^{-3/2}\right)
\end{aligned}
$$

for $0<\alpha+\delta<1$, where $P^{**}$ is the bootstrap distribution conditional on $\mathcal{X}^*$, the third equality follows from the bootstrap version of the Cornish Fisher expansion (4.14) on the quantile $Q_s^{**}(\cdot)$ of the distribution of $\sqrt{n} \cdot \widehat{\overline{\mathrm{IF}}}^*\left(\mathcal{X}^{**}\right)/\hat{\sigma}^*$ given $\mathcal{X}^*$ along with a Taylor expansion, such that for any $0<\beta<1$,

$$
Q_s^{**}\left(\beta+O_{P^*}\left(n^{-3/2}\right)\right)=Q_s^{**}(\beta)+O_{P^*}\left(n^{-3/2}\right),
$$

and the forth equality is a direct result by the delta method for Edgeworth expansions.

Remember $\tilde{\delta}_n$ is defined in (4.11) so that $\tilde{\pi}(\alpha + \tilde{\delta}_n) - (1 - \alpha) = 0$. As with (4.27), we deduce

$$\begin{aligned}\tilde{\delta}_n &= -\hat{c}_0 \hat{d}_n^{-1} n^{-1} + O_P\left(n^{-3/2}\right) \\ &= \delta_n - n^{-3/2}\hat{\Delta}_n + O_P\left(n^{-3/2}\right) \\ &= \delta_n + O_P\left(n^{-3/2}\right).\end{aligned}$$

Therefore, using the bootstrap Cornish-Fisher expansion (4.17) and Taylor series expansion, we have

$$Q_s^*\left(\beta_1 \pm \beta_2 \tilde{\delta}_n\right) = Q_s^*\left(\beta_1 \pm \beta_2 \delta_n\right) + O_P\left(n^{-3/2}\right) \tag{4.28}$$

for any $0 < \beta_1, \beta_2 < 1$.

Our calibrated fast double bootstrap confidence interval has the form:

$$I_0(\alpha + \tilde{\delta}_n; \mathcal{X}, \mathcal{X}^*) = \left[\hat{\mu} - Q^*(1 - (\alpha + \tilde{\delta}_n)/2), \hat{\mu} - Q^*((\alpha + \tilde{\delta}_n)/2)\right]$$

Now, we can show third order accuracy as follows:

$$
\begin{aligned}
&\pi(\alpha+\tilde{\delta}_n)-(1-\alpha)\\
&=P\left[\mu^{\dagger}\in I_0(\alpha+\tilde{\delta}_n;\mathcal{X},\mathcal{X}^*)\right]-(1-\alpha)\\
&=P\left[\hat{\mu}-Q^*\left(1-\frac{\alpha+\tilde{\delta}_n}{2}\right)\leq\mu^{\dagger}\leq\hat{\mu}-Q^*\left(\frac{\alpha+\tilde{\delta}_n}{2}\right)\right]-(1-\alpha)\\
&=P\left[Q_s^*\left(\frac{\alpha+\tilde{\delta}_n}{2}\right)\leq\frac{\sqrt{n}\left(\hat{\mu}-\mu^{\dagger}\right)}{\hat{\sigma}}\leq Q_s^*\left(1-\frac{\alpha+\tilde{\delta}_n}{2}\right)\right]-(1-\alpha)\\
&=P\left[Q_s^*\left(\frac{\alpha+\delta_n}{2}\right)+O_P\left(n^{-3/2}\right)\leq\frac{\sqrt{n}\left(\hat{\mu}-\mu^{\dagger}\right)}{\hat{\sigma}}\leq Q_s^*\left(1-\frac{\alpha+\delta_n}{2}\right)+O_P\left(n^{-3/2}\right)\right]\\
&\qquad-(1-\alpha)\\
&=P\left[Q_s^*\left(\frac{\alpha+\delta_n}{2}\right)\leq\frac{\sqrt{n}\left(\hat{\mu}-\mu^{\dagger}\right)}{\hat{\sigma}}\leq Q_s^*\left(1-\frac{\alpha+\delta_n}{2}\right)\right]-(1-\alpha)+O\left(n^{-3/2}\right)\\
&=\pi(\alpha+\delta_n)-(1-\alpha)+O\left(n^{-3/2}\right)\\
&=O\left(n^{-3/2}\right),
\end{aligned}
$$

where the fourth equality follows by (4.28), the fifth equality follows by the delta method of Edgeworth expansions and the last equality follows because $\pi(\alpha+\delta_n)-(1-\alpha)=0$.

## 4.4 Simulation Study

We compared the finite sample coverage of confidence intervals computed using our proposed fast double bootstrap method and other existing procedures in a simulation study.

We focused on estimating the causal effect of a treatment on an outcome from observational data with the inverse probability weighted (IPW) estimator. We define $Y_1$ and $Y_0$ to be potential outcomes under treatment 1 and 0, respectively. In the observational study, we assume that we observe $n$ independent and identically distributed (i.i.d.) copies of $X = (Z, T, Y)'$, where $Z$ is the covariate, $T$ is binary treatment assignment indicator and $Y = TY_1 + (1-T)Y_0$ is the outcome. The goal is to draw inference about $\mu_t^\dagger = E[Y_t]$ $(t = 0, 1)$ and $\mu_1^\dagger - \mu_0^\dagger$.

To identify $\mu_t^\dagger$, we assume that $T$ is independent of $(Y_1, Y_0)$ given $Z$. To estimate $\mu_t^\dagger$, we considered the widely used inverse probability weighted (IPW) estimation approach. In this approach, a logistic regression model is specified for $P[T = t|Z]$, i.e.,

$$P\left[T = t|\, Z\right] = \frac{\exp\left\{tl(Z;\gamma^\dagger)\right\}}{1 + \exp\left\{l(Z;\gamma^\dagger)\right\}}.$$

where $l(Z;\gamma)$ is a specified function of $Z$ and $\gamma$. Under the identification and modeling assumptions, it follows that

$$\psi_t(X;\mu_t,\gamma) = \frac{I\left\{T = t\right\}}{\omega_t(Z;\gamma)}\left(Y - \mu_t\right), \tag{4.29}$$

is an unbiased estimating function, where

$$\omega_t(Z;\gamma) = \frac{\exp\left\{tl(Z;\gamma)\right\}}{1 + \exp\left\{l(Z;\gamma)\right\}}$$

That is, $E[\psi_t(X;{\mu_t}^\dagger,\gamma^\dagger)] = 0$.

Given data on a sample $\mathcal{X} = (X_1, \ldots, X_n)$ where $X_i = (Z_i, T_i, Y_i)'$ $(i = 1, \ldots, n)$, the nuisance parameter $\gamma$ can be estimated by solving the score equation of the logistic regression of $T$ given $Z$, i.e. $\hat{\gamma}$ defined as the solution to

$$\frac{1}{n}\sum_{i=1}^{n} S_\gamma(T_i, Z_i;\gamma) = 0,$$

where the score function

$$S_\gamma(T, Z; \gamma) = \frac{\partial l(Z;\gamma)}{\partial \gamma} \{T - \omega_1(Z;\gamma)\}.$$

Then the estimator $\hat{\mu}_t$ can be obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n} \psi_t(X_i; \mu_t, \hat{\gamma}) = 0.$$

It can be shown that $\hat{\mu}_t$ is a regular and asymptotically linear (RAL) estimator of $\mu_t^\dagger$, with the influence function

$$\mathrm{IF}_t(X; \mu_t^\dagger, \gamma^\dagger) = \psi_t(X; \mu_t^\dagger, \gamma^\dagger) - E\left[\frac{\partial \psi_t(X; \mu_t^\dagger, \gamma^\dagger)}{\partial \gamma'}\right] E\left[\frac{\partial S_\gamma(T, Z; \gamma^\dagger)}{\partial \gamma'}\right]^{-1} S_\gamma(T, Z; \gamma^\dagger).$$

The causal effect $\mu_1^\dagger - \mu_0^\dagger$ can be estimated by $\hat{\mu}_1 - \hat{\mu}_0$ and will have influence function:

$$\mathrm{IF}_1(X; \mu_1^\dagger, \gamma^\dagger) - \mathrm{IF}_0(X; \mu_0^\dagger, \gamma^\dagger).$$

In our simulation study, we generated data as follows:

- $Z$ is drawn from a mixture distribution where

$$Z \sim \begin{cases} Unif(-1, 1) & \text{with probability } 0.8 \\ 1 & \text{with probability } 0.2 \end{cases}$$

- Given $Z$, the three variables $T$, $Y_1$ and $Y_0$ are independently distributed as

$$T \sim \text{Bernoulli}\left(\text{expit}\left\{-3.5Z\right\}\right)$$

$$Y_1 \sim \mathcal{N}\left(Z,\ 0.0016\right)$$

$$Y_0 \sim \mathcal{N}\left(1 + 1.5Z,\ 0.0016\right)$$

Under the above settings, $\mu_1^\dagger = 0.2,\ \mu_0^\dagger = 1.3$ and $\mu_1^\dagger - \mu_0^\dagger = -1.1$. Notice that the model $P[T=1|Z]$ was correctly specified in such way, with the true value $\gamma^\dagger = (0, -3.5)'$. Under these assumptions, $P[T=1] = 0.406$.

We considered the following five types of $(1-\alpha)$-level equal-tail two-sided confidence intervals for each parameter of interest $\mu$.

- Influence function based Wald type (IF): $\left[\hat{\mu} - n^{-1/2}\hat{\sigma}z_{1-\alpha/2}, \hat{\mu} - n^{-1/2}\hat{\sigma}z_{\alpha/2}\right]$;
- Single bootstrap percentile (SB-p): $\left[\tilde{Q}^*(\alpha/2), \tilde{Q}^*(1-\alpha/2)\right]$, where $\tilde{Q}^*(x)$ is the $x$-level quantile of conditional distribution of $\hat{\mu}^*$ given $\mathcal{X}$;
- Single bootstrap-t (SB-t): $\left[\hat{\mu} - n^{-1/2}\hat{\sigma}\tilde{Q}_s^*(1-\alpha/2), \hat{\mu} - n^{-1/2}\hat{\sigma}\tilde{Q}_s^*(\alpha/2)\right]$, where $\tilde{Q}_s^*(x)$ is the $x$-level quantile of the conditional distribution of $\sqrt{n}(\hat{\mu}^* - \hat{\mu})/\hat{\sigma}^*$ given $\mathcal{X}$;
- Influence function based single bootstrap (SB-IF): $\left[\hat{\mu} - Q^*(1-\alpha/2), \hat{\mu} - Q^*(\alpha/2)\right]$, where $Q^*(x)$ is the $x$-level quantile of the conditional distribution of $\widehat{\overline{\text{IF}}}(\mathcal{X}^*)$ given $\mathcal{X}$;
- Influence function based fast double bootstrap (FDB): $\left[\hat{\mu} - Q^*(1-(\alpha+\tilde{\delta}_n)/2), \hat{\mu} - Q^*((\alpha+\tilde{\delta}_n)/2)\right]$, where $Q^*(x)$ is the $x$-level quantile of the conditional distribution of $\widehat{\overline{\text{IF}}}(\mathcal{X}^*)$ given $\mathcal{X}$, and $\tilde{\delta}_n$ is defined in (4.11).

In practice, the theoretical quantiles and probabilities needed in computing the confidence intervals above were approximated by the corresponding empirical estimates, based on the $B$ resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$ from $\mathcal{X}$. There is no need for any inner-level resamplings. Only the fast double bootstrap has third order accuracy; the remaining procedures are second order accurate.

To illustrate the coverage accuracy and its convergence speed (to nominal level) for various confidence intervals under different choices of sample size, especially for small to medium samples, we set the sample size $n = 200, 500$ and 1000 successively, and for each of the three sample sizes we simulated 5000 data sets. Given each data set $\mathcal{X} = (X_1, \ldots, X_n)$, $B = 5000$ non-parametric multinomial bootstrap resamples were drawn, i.e. $\mathcal{X}^* = (X_1^*, \ldots, X_n^*)$ i.i.d. from the empirical distribution $\hat{F}_n(x) = n^{-1} \sum_{i=1}^{n} 1\{X_i \leq x\}$. We then constructed 95% and 90% equal-tail two-sided confidence intervals using each of the five procedures accordingly, for each of the three estimands of interest $\mu_1^\dagger, \mu_0^\dagger$ and $\mu_1^\dagger - \mu_0^\dagger$. Finally, the empirical coverage of each choice of confidence interval (over possible procedures, nominal levels, and target parameters) was computed across the 5000 simulated data sets, for each sample size. The results are presented in Table 4.1.

It is important to notice that by the simulation design, the treatment assignment probability $\omega_t(Z; \gamma^\dagger) = P[T = t | Z]$ has very small value when the covariate $Z$ approaches its minimum at -1 or maximum at 1:

$$P[T = 1 | Z = 1] = P[T = 0 | Z = -1] = \text{expit}(-3.5) = 0.029,$$

which makes the inverse probability weight large for certain observations in the IPW estimating function (4.29). It is well known that large weights can lead to unstable performance of the IPW estimator especially when the sample size is insufficient for the asymptotic normal approximation to work satisfactorily. We expect this to happen for both treatment arms in our simulation.

In general, Table 4.1 lends support to the higher order coverage accuracy of our fast double bootstrap confidence interval procedure. For both 95% and

Table 4.1: Comparison of empirical coverage probabilities for various equal-tail two-sided confidence interval (CI) procedures, with increasing sample sizes

| Sample size | Procedure[a] | 95% CI coverage (%) | | | 90% CI coverage (%) | | |
|---|---|---|---|---|---|---|---|
| | | $\mu_1$ | $\mu_0$ | $\mu_1 - \mu_0$ | $\mu_1$ | $\mu_0$ | $\mu_1 - \mu_0$ |
| 200 | IF | 81.5 | 82.3 | 84.2 | 77.6 | 77.5 | 79.2 |
| | SB-p | 83.5 | 85.0 | 83.7 | 79.6 | 79.9 | 78.5 |
| | SB-t | 86.0 | 86.8 | 87.5 | 81.8 | 81.0 | 82.6 |
| | SB-IF | 79.7 | 81.1 | 82.1 | 75.8 | 76.2 | 78.0 |
| | FDB | 85.3 | 88.1 | 90.0 | 82.2 | 82.9 | 85.9 |
| 500 | IF | 91.8 | 89.5 | 90.8 | 86.4 | 84.3 | 85.0 |
| | SB-p | 92.9 | 90.6 | 90.4 | 87.7 | 85.3 | 84.6 |
| | SB-t | 94.3 | 91.9 | 93.0 | 91.1 | 86.9 | 88.2 |
| | SB-IF | 89.5 | 88.1 | 89.0 | 84.9 | 83.4 | 83.7 |
| | FDB | 94.7 | 92.5 | 94.5 | 90.7 | 87.6 | 89.4 |
| 1000 | IF | 93.2 | 92.2 | 92.7 | 88.1 | 87.5 | 88.2 |
| | SB-p | 94.1 | 92.8 | 92.8 | 88.9 | 88.1 | 88.1 |
| | SB-t | 95.2 | 94.5 | 93.6 | 90.7 | 89.8 | 88.5 |
| | SB-IF | 92.3 | 91.0 | 91.8 | 87.2 | 86.9 | 87.6 |
| | FDB | 95.5 | 94.6 | 94.6 | 90.2 | 89.5 | 89.9 |

[a] Procedures: influence function based Wald type (IF), single bootstrap percentile (SB-p), single bootstrap-t (SB-t), influence function based single bootstrap (SB-IF), influence function based fast double bootstrap (FDB).

90% nominal levels, the fast double bootstrap intervals tend to have smaller coverage error than the other four types of intervals. The advantage is most noticeable for estimating $\mu_1^\dagger$ and $\mu_1^\dagger - \mu_0^\dagger$. Though all five procedures yield coverage lower than nominal levels at sample size 200, the fast double bootstrap interval coverage approach the nominal levels for $\mu_1^\dagger$ and $\mu_1^\dagger - \mu_0^\dagger$ by sample size 500, and for $\mu_0^\dagger$ by sample size 1000. Among the other four procedures, the SB-t intervals outperform the other procedures. The influence-function single bootstrap performed poorly, which emphasizes the need for calibration.

## 4.5 Conclusion and discussion

In this paper, we proposed the fast double bootstrap confidence interval based on influence function resampling and saddlepoint approximation methods. Our procedure yields equal-tail two-sided confidence intervals of third order accuracy, without inner-level resamplings.

The simulations indicate that our fast double bootstrap intervals are more reliable than other widely used intervals like Wald type, bootstrap percentile or bootstrap-t for small to medium sample size where asymptotic normality fails to provide a reasonable approximation. Given the same computational burden as most single layer bootstrap methods, our technique is economical and easy to apply, while full implementation of the double bootstrap procedure can be prohibitive.

For simplicity, we have assumed the parameter of interest is scalar. However, our approach is readily generalizable to vector valued parameters. Our investigation of coverage accuracy relied on Edgeworth and Cornish-Fisher expansions, which requires the nuisance estimator $\hat{\gamma}$ to be $\sqrt{n}$-consistent. This is typically true for parametric models. However, in most semi-parametric models where the nuisance parameter is estimated using non-parametric methods or complex (e.g. sequential) procedures, it may only be estimable at rates slower than $\sqrt{n}$. An important direction of future research is higher order asymptotics and bootstrap convergence rate of semi-parametric estimations when the nuisance parameter is not $\sqrt{n}$-estimable.

# Chapter 5

# Conclusion

In this dissertation, we developed statistical methodology using influence functions, under various sampling designs. We demonstrated the power of influence functions as an inferential tool:

- In causal inference for comprehensive cohort studies, we found efficient estimators under certain modeling restrictions by investigating the geometry of the class of influence functions. We derived the properties of the resulting estimators, like robustness to model misspecification, through investigation of the corresponding influence function.

- In optimal outcome-dependent two-phase sampling design, the asymptotic variance of the estimators under discussion, as the target to be minimized, was accessed using their influence functions. The problem was tackled by identifying the key component in the influence function which determines the relationship between the sampling fractions of interest and the asymptotic variance of the resulting estimator.

- In the fast double bootstrap procedure, we replaced the pivot of the original estimator with its asymptotic linear representation of the influence

function in the resampling process. This shows how the influence function can be used to approximate the asymptotic behavior of the estimator under discussion. Furthermore, the representation of the estimator as an average of independent and identically distributed influence functions allows us to apply many of the existing methods (e.g. saddlepoint approximation), and therefore is more tractable compared to using the original estimator.

In summary, statistical inference based on influence functions is widely applicable in biomedical and public health research. The topics discussed in this dissertation suggest enormous power and potential of the the influence functions as inferential tools, which can be further exploited in future research.

# Bibliography

Armstrong, T. B., Bertanha, M., and Hong, H. (2014). A fast resample method for parametric and semiparametric models. *Journal of Econometrics*, 179(2):128–133.

Baillargeon, S. and Rivest, L.-P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77(3):331–344.

BARI Investigators (1996). Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. *New England Journal of Medicine*, 335(4):217–225.

Bedi, N., Lee, A., Harrison, G., Chilvers, C., Dewey, M., Fielding, K., Miller, P., Gretton, V., Williams, I., Churchill, R., et al. (2000). Assessing effectiveness of treatment of depression in primary care. *The British Journal of Psychiatry*, 177(4):312–318.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468.

Breslow, N. E. (2000). Statistics in the life and medical sciences. *Journal of the American Statistical Association*, 95(449):281–282.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468.

Breslow, N. E. and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):447–461.

Breslow, N. E. and Holubkov, R. (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*, 16(1-3):103–116.

Breslow, N. E., Wellner, J. A., and McNeney, B. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Annals of Statistics*, 31(4):1110–1139.

Brewin, C. R. and Bradley, C. (1989). Patient preferences and randomised clinical trials. *BMJ: British Medical Journal*, 299(6694):313–315.

Brocklehurst, P. (1997). Partially randomised patient preference trials. *BJOG: An International Journal of Obstetrics & Gynaecology*, 104(12):1332–1335.

Brooks, M. M., Jones, R. H., Bach, R. G., Chaitman, B. R., Kern, M. J., Orszulak, T. A., Follmann, D., Sopko, G., Blackstone, E. H., Califf, R. M., et al. (2000). Predictors of mortality and mortality from cardiac causes in

the Bypass Angioplasty Revascularization Investigation (BARI) randomized trial and registry. *Circulation*, 101(23):2682–2689.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):573–585.

Chang, J. and Hall, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika*, pages 203–214.

Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.

Cochran, W. G. (1963). *Sampling Techniques*. John Wiley & Sons.

Colantuoni, E. and Rosenblum, M. (2015). Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine*, 34(18):2602–2617.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA: the Journal of the American Medical Association*, 276(11):889–897.

Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica*, 49(5):1289–1316.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, 51(3):765–782.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, pages 631–650.

Daniels, H. E. (1987). Tail probability approximations. *International Statistical Review/Revue Internationale de Statistique*, pages 37–48.

Davidson, R. and MacKinnon, J. G. (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics & Data Analysis*, 51(7):3259–3281.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge University Press.

Detre, K. M., Guo, P., Holubkov, R., Califf, R. M., Sopko, G., Bach, R., Brooks, M. M., Bourassa, M. G., Shemin, R. J., Rosen, A. D., et al. (1999). Coronary revascularization in diabetic patients a comparison of the randomized and observational components of the Bypass Angioplasty Revascularization Investigation (BARI). *Circulation*, 99(5):633–640.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, pages 189–212.

Diciccio, T. J. and Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to bayesian and conditional inference. *Biometrika*, 78(4):891–902.

DiCiccio, T. J., Martin, M. A., and Young, G. A. (1992). Fast and accurate approximate double bootstrap confidence intervals. *Biometrika*, 79(2):285–295.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Fears, T. R. and Brown, C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42(4):955–960.

Feit, F., Brooks, M. M., Sopko, G., Keller, N. M., Rosen, A., Krone, R., Berger, P. B., Shemin, R., Attubato, M. J., Williams, D. O., et al. (2000). Long-term clinical outcome in the Bypass Angioplasty Revascularization Investigation registry comparison with the randomized trial. *Circulation*, 101(24):2795–2802.

Fielding, L. P., Grace, R., and Hittinger, R. (1999). Patients who are eligible but not randomised should be included as additional comparative arm in study. *BMJ: British Medical Journal*, 318(7187):874.

Giacomini, R., Politis, D. N., and White, H. (2013). A warp-speed method for conducting monte carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(03):567–589.

Gilbert, P. B., Yu, X., and Rotnitzky, A. (2014). Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in Medicine*, 33(6):901–917.

Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, pages 1431–1452.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hall, P. and Maesono, Y. (2000). A weighted bootstrap approach to bootstrap iteration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):137–144.

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika*, 75(4):661–671.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Henshaw, R., Naji, S., Russell, I., and Templeton, A. (1993). Comparison of medical abortion with surgical vacuum aspiration: women's preferences and acceptability of treatment. *BMJ: British Medical Journal*, 307(6906):714–717.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278.

Jensen, L., Vestergaard, P., Hermann, A., Gram, J., Eiken, P., Abrahamsen, B., Brot, C., Kolthoff, N., Sørensen, O., Beck-Nielsen, H., et al. (2003). Hormone replacement therapy dissociates fat mass and bone mass, and tends to reduce weight gain in early postmenopausal women: A randomized controlled 5-year

clinical trial of the danish osteoporosis prevention study. *Journal of bone and mineral research*, 18(2):333–342.

Kerry, S., Hilton, S., Patel, S., Dundas, D., Rink, E., and Lord, J. (2000). Routine referral for radiography of patients presenting with low back pain: is patients' outcome influenced by gps' referral for plain radiography? *Health technology assessment (Winchester, England)*, 4(20):1–119.

King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M., Morou, M., Sibbald, B., and Lai, R. (2005). Impact of participant and physician intervention preferences on randomized trials: a systematic review. *Jama*, 293(9):1089–1099.

King, M., Sibbald, B., Ward, E., Bower, P., Lloyd, M., Gabbay, M., and Byford, S. (2000). Randomised controlled trial of non-directive counselling, cognitive-behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care. *Health technology assessment (Winchester, England)*, 4(19):1–83.

King, S. B., Barnhart, H. X., Kosinski, A. S., Weintraub, W. S., Lembo, N. J., Petersen, J. Y., Douglas, J. S., Jones, E. L., Craver, J. M., Guyton, R. A., et al. (1997). Angioplasty or surgery for multivessel coronary artery disease: comparison of eligible registry and randomized patients in the east trial and influence of treatment selection on outcomes. *The American journal of cardiology*, 79(11):1453–1459.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of*

*the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438.

Marcus, S. M. (1997). Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *Journal of clinical epidemiology*, 50(7):823–828.

Nankervis, J. C. (2005). Computational algorithms for double bootstrap confidence intervals. *Computational Statistics & Data Analysis*, 49(2):461–475.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116.

Nicolaides, K., Brizot, M., Patel, F., and Snijders, R. (1994). Comparison of chorionic villus sampling and amniocentesis for fetal karyotyping at 10-13 weeks' gestation. *The Lancet*, 344(8920):435–439.

Olschewski, M. and Scheurlen, H. (1985). Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial. *Methods of Information in Medicine*, 24:131–134.

Olschewski, M., Schumacher, M., and Davis, K. B. (1992). Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled clinical trials*, 13(3):226–239.

Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413):108–113.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society. Series B. Methodological*, 57(2):409–424.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

Rovers, M. M., Straatman, H., Ingels, K., van der Wilt, G.-J., van den Broek, P., and Zielhuis, G. A. (2001). Generalizability of trial results based on randomized versus nonrandomized allocation of ome infants to ventilation tubes or watchful waiting. *Journal of clinical epidemiology*, 54(8):789–794.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Rücker, G. (1989). A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Statistics in medicine*, 8(4):477–485.

Schill, W., Jöckel, K. H., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika*, 80(2):339–352.

Schmoor, C., Caputo, A., and Schumacher, M. (2008). Evidence from nonrandomized studies: a case study on the estimation of causal effects. *American journal of epidemiology*, 167(9):1120–1129.

Schmoor, C., Olschewski, M., and Schumacher, M. (1996). Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in medicine*, 15(3):263–271.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47(2):497–510.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Torgerson, D. J. and Sibbald, B. (1998). Understanding controlled trials. what is a patient preference trial? *BMJ: British Medical Journal*, 316(7128):360.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Tu, D. and Shao, J. (1995). *The Jackknife and bootstrap*. Springer Series in Statistics, New York.

Wang, W., Scharfstein, D., Tan, Z., and MacKenzie, E. J. (2009). Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):947–969.

Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469.

Wennberg, J. E., Barry, M. J., Fowler, F. J., and Mulley, A. (1993). Outcomes research, ports, and health care reform. *Annals of the New York Academy of Sciences*, 703(1):52–62.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128.

# Yi Lu

**Johns Hopkins University, Department of Biostatistics**

**615 N. Wolfe Street, E3037, Baltimore, MD 21205, USA**

**Phone: 410-502-3365** **Email: ylu18@jhu.edu**

## Education

*Ph.D. in Biostatistics*, Johns Hopkins University, 2016

Advisor: Dr. Daniel O. Scharfstein

*B.S. in Mathematics*, Fudan University, China, 2009

## Research Experience

2014 – 2016 **Research Assistant** to Dr. Daniel O. Scharfstein

Johns Hopkins University School of Public Health

- Project: comprehensive cohort analysis of FIXIT study, Major Extremity Trauma Research Consortium (Principal Investigator: Dr. Ellen MacKenzie)
- Project: fast double bootstrap confidence intervals

2011 – 2014 **Research Assistant** to Dr. Mei-Cheng Wang

Johns Hopkins University School of Public Health

- Project: survival and longitudinal analyses of cognitive decline, BIOCARD study (Principal Investigator: Dr. Marilyn S. Albert)

- Project: power of within-center randomization designs with correlated binary data (Principal Investigator: Dr. Lawrence Wissow)

2009 **Research Internship**

Bioinformatics Lab, School of Life Sciences, Fudan University, Shanghai, China

- Project: statistical analysis of human biological markers polymorphism in Indian population study
- Supervisor: Dr. Fuli Yu

## Teaching Experience

2011 – 2015 **Teaching Assistant**

Department of Biostatistics, Johns Hopkins University School of Public Health

- Design of Clinical Experiments, 2015.
- Statistical Methods in Public Health I – II, 2014.
- Statistical Methods in Public Health III – IV, 2013 – 2014.
- Survival Analysis II, 2012.
- Survival Analysis I, 2012 – 2013.
- Statistical Methods in Public Health I – IV, 2011 – 2012.

## Honors and Awards

2010 – 2016 Department of Biostatistics Graduate Scholarship,

*Johns Hopkins University*

2009 B.S. with First Class Honors, *Fudan University*

2006 – 2009 People's Scholarship, *Fudan University*

## Publications

**Lu, Y.**, Scharfstein, D. O. (2016). Causal inference for comprehensive cohort studies. *In preparation.*

**Lu, Y.**, Scharfstein, D. O. (2016). Influence function based fast double bootstrap confidence intervals. *In preparation.*

**Lu, Y.**, Scharfstein, D. O. (2016). Optimal outcome-dependent two-phase sampling. *In preparation.*

Soldana, A., Pettigrewa, C., **Lu, Y.**, Wang, M-C., Selnes, O., Albert, M., Brown, T., Ratnanather, T., Younes, L., Miller, M. I., and the BIOCARD Research Team. (2015). Relationship of medial temporal lobe atrophy, APOE genotype, and cognitive reserve in preclinical Alzheimers disease. *Human brain mapping*, 36(7): 2826–2841.

Albert, M., Soldan, A., Gottesman, R., McKhann, G., Sacktor, N., Farrington, L., Grega, M., Turner, RS., **Lu, Y.**, Li, S., Wang, M-C., Selnes, O. and the BIOCARD Research Team. (2014). Cognitive changes preceding clinical symptom onset of mild cognitive impairment and relationship to ApoE genotype. *Current Alzheimer research*, 11(8): 773–784.

Moghekar, A., Li, S., **Lu, Y.**, Li, M., Wang, M-C., Albert, M., O'Brien, R. and the BIOCARD Research Team. (2013). Cerebrospinal Fluid

Biomarker Changes Precede Symptom Onset Mild Cognitive Impairment. *Neurology*, 81(20):1753–1758.

Pettigrew, C., Soldan, A., Li, S., **Lu, Y.**, Wang, M-C., Selnes, O., Moghekar, A., O'Brien, R., Albert, M. and the BIOCARD Research Team. (2013). Relationship of Cognitive Reserve and APOE Status to the Emergence of Clinical Symptoms in Preclinical Alzheimers Disease. *Cognitive Neuroscience*, 4(3-4): 136–142.

## Presentations

"Cortical regions are associated with risk of clinical symptom onset during preclinical Alzheimers disease", Annual Meeting of the Society for Neuroscience, Washington, DC, Nov. 2014, *poster, contributed*.

"Mixed Effects Models for Analyzing the Association of CSF at Baseline and Change in Cognition over Time", BIOCARD Study Scientific Advisory Board Meeting, Baltimore, Mar. 2014.

"Inferential Approaches to Relative Risk Regression", Eastern North American Region Meetings, Baltimore, Mar. 2014, *poster*.

"Inferential Approaches to Relative Risk Regression", Causal Inference Group Meeting, Department of Biostatistics, Johns Hopkins University, Baltimore, Feb. 2014.

"Introduction to BIOCARD ('Biomarkers of Cognitive Decline Among Normal Individuals: the BIOCARD cohort') – Biostatistics Core", Student Journal Club, Department of Biostatistics, Johns Hopkins University,

Baltimore, Mar. 2012.

"Estimation of Change-Points in Hazard Rates for Nonparametric/Cox Regression Models", Survival, Longitudinal and Multivariate (SLAM) Data Working Group Meeting, Department of Biostatistics, Johns Hopkins University, Baltimore, Jan. 2012.

## Professional Memberships

2013 – present Eastern North American Region (ENAR)

International Biometric Society

## Skills

Programming: R, MATLAB, STATA, C/C++

Computer Application: Microsoft Office, LaTeX

Language: English, Chinese-Mandarin